Multi-level methods for degenerated problems with applications to p-versions of the fem

Von der Fakultät für Mathematik der Technischen Universität Chemnitz

genehmigte

D i s s e r t a t i o n

zur Erlangung des akademischen Grades

Doctor rerum naturalium

(Dr. rer. nat.)

vorgelegt

von Dipl.-Math. Sven Beuchler

geboren am 24. Juli 1975 in Karl-Marx-Stadt

eingereicht am 5. 2. 2003

Gutachter: Prof. Dr. Arnd Meyer
Prof. Dr. Arnold Reusken
PD Dr. Markus Melenk

Tag der Verteidigung: 11. 7. 2003

Preface

Many physical problems lead to boundary value problems for partial differential equations which can be solved with the $h-$, $hp-$, and $p-$version of the finite element method. Such a discretization leads to a system of linear algebraic equations. One of the most efficient methods in order to solve systems of linear algebraic equations resulting from p-version finite element discretizations of elliptic boundary value problems is the conjugate gradient method with domain decomposition preconditioners. The ingredients of such a preconditioner are a preconditioner for the Schur complement, a preconditioner related to the Dirichlet problems in the sub-domains, and an extension operator from the boundaries of the sub-domains into their interior.
The aim of this monograph is to develop a preconditioner for the problems in the sub-domains. For the Poisson equation, the preconditioner for this problem can be interpreted as the stiffness matrix resulting from an h-version finite element discretization of a degenerated operator. The corresponding systems of finite element equations are solved by a multi-grid algorithm. Alternatively, a preconditioned conjugate gradient method is used, where the preconditioner is a multi-grid preconditioner, an AMLI preconditioner, or a so-called MTS-BPX preconditioner. A rigorous mathematical theory analyzing the condition numbers of the preconditioned systems and the convergence rate of the multi-grid algorithm is given. The analysis is purely algebraic and basically relies on two ingredients, the strengthened Cauchy-inequality and the construction of the smoother.
This work has been possible only with the help, stimulation and encouragement of many people. I want to thank Prof. Arnd Meyer for the supervision of my dissertation. Furthermore, I wish to express my particular appreciation to Dr. Michael Jung for many stimulations, fruitful discussions and proofreading. Chapter 7 comprises the results of a joint work with Prof. Reinhold Schneider and Prof. Christoph Schwab. I would like to thank both for their contributions and ideas. Furthermore, I would like to thank all colleagues of the faculty of mathematics at the TU-Chemnitz for the stimulating working atmosphere. Special thanks go to Dr. Gerd Kunert for improving the English and Roman Unger for removing all my LaTeX problems. This work was supported by the Deutsche Forschungsgemeinschaft. At last I would like to thank my father for his support and patience over the years. All this help and support is gratefully acknowledged.

List of symbols

In this section, a list of the most important symbols is given.

- Domains:
  d - space dimension,
  $I = (0,1)$,
  $\Omega = (0,1)^2$,
  $\Omega_3 = (0,1)^3$,
  $\mathcal{R}_d = (-1,1)^d$.

- Bilinear forms:
  $a_\triangle(u,v) = \int u_x v_x + u_y v_y$,
  $a_s(u,v) = \int_0^1 u'(x)v'(x)\,\mathrm{d}x$,
  $a_m(u,v) = \int_0^1 x^2 u(x)v(x)\,\mathrm{d}x$,
  $a_{\overline{m}}(u,v) = \int_0^1 x^{-2} u(x)v(x)\,\mathrm{d}x$,
  $a_1(u,v) = a_s(u,v) + a_m(u,v) + a_{\overline{m}}(u,v)$,
  $a(u,v) = \int_\Omega y^2 u_x v_x + x^2 u_y v_y$,
  $a_3(u,v) = \int_{\Omega_3} x^2 u_{yz} v_{yz} + y^2 u_{xz} v_{xz} + z^2 u_{xy} v_{xy}$.

- Polynomials:
  p - polynomial degree,
  L_i - i-th Legendre polynomial,
  $\hat{L}_i$ - i-th integrated Legendre polynomial,
  T_i - i-th Chebyshev polynomial.

- Mesh parameter and shape functions:
  k - level number,
  $n = 2^k$,
  τ_i^k - interval $\left(\frac{i}{n}, \frac{i+1}{n}\right)$,
  x_{ij}^k - node $\frac{1}{n}(i,j)$,
  $\tau_{ij}^{1,k}$ - triangle with vertices x_{ij}^k, $x_{i,j+1}^k$ and $x_{i+1,j+1}^k$,
  $\tau_{ij}^{2,k}$ - triangle with vertices x_{ij}^k, $x_{i+1,j}^k$ and $x_{i+1,j+1}^k$,
  $\mathcal{E}_{ij}^k$ - square $\overline{\tau_i^k \times \tau_j^k}$,
  $\mathcal{H}_{ijl}^k$ - cube $\overline{\tau_i^k \times \tau_j^k \times \tau_l^k}$,
  $\phi_i^{(1,k)}$ - piecewise linear nodal hat function with $\phi_i^{(1,k)}(\frac{j}{n}) = \delta_{ij}$,
  ϕ_{ij}^k - piecewise linear nodal hat function with $\phi_{ij}^k(x_{lm}^k) = \delta_{il}\delta_{jm}$,
  $\phi_{b,ij}^k$ - piecewise bilinear nodal hat function with $\phi_{b,ij}^k(x_{lm}^k) = \delta_{il}\delta_{jm}$,
  $\phi_{t,ijl}^k(x,y,z) = \phi_i^{(1,k)}(x)\phi_j^{(1,k)}(y)\phi_l^{(1,k)}(z)$.

- Norms and function spaces:

$L^2(\Omega)$	-	$\{u : \Omega \mapsto \mathbb{R},\ u$ measurable, $\int_\Omega u^2\ \mathrm{d}x < \infty\}$,
$H^1(\Omega)$	-	$\{u \in L^2(\Omega), \nabla u \in (L^2(\Omega))^d\}$, $\Omega \subset \mathbb{R}^d$
$H_0^1(\Omega)$	-	$\{u \in H^1(\Omega), u = 0$ on $\partial\Omega\}$,
$\omega(\xi)$	-	weight function,
$L^2_\omega((a,b))$	-	$\{u \in L^2((a,b)), \int_a^b \omega^2(x)u^2(x)\ \mathrm{d}x < \infty\}$,
$\|\cdot\|_0$	-	L^2-norm,
$\|\cdot\|_1$	-	H^1-norm,
$\|\cdot\|_\omega$	-	L^2_ω-norm,
$\|\cdot\|_a$	-	energetic-norm,
$\|\cdot\|_F$	-	Frobenius norm of a matrix,

- quadratic matrices:

$\lambda_{min}(M)$	-	smallest eigenvalue of M,
$\lambda_{max}(M)$	-	largest eigenvalue of M,
$\kappa(M)$	-	condition number of M in 2-norm,
$\kappa\left(A^{-1}B\right)$	-	condition number of $A^{-1/2}BA^{-1/2}$, if A and B are symmetric and positive definite,
$\det(M)$	-	determinant of M,
$\mathrm{trace}(M)$	-	trace of M,
$\mathrm{diag}[\mathfrak{a}]$	-	diagonal matrix with the main diagonal equal to the vector $\mathfrak{a}$,
$\mathrm{tridiag}[\mathfrak{a},\mathfrak{b}]$	-	tridiagonal symmetric matrix with main diagonal $\mathfrak{a}$ and first sub-diagonal $\mathfrak{b}$,
$\mathrm{pentdiag}[\mathfrak{a},\mathfrak{b},\mathfrak{c}]$	-	penta-diagonal symmetric matrix with main diagonal $\mathfrak{a}$ and sub-diagonals $\mathfrak{b}$ and $\mathfrak{c}$,
$\mathrm{blockdiag}\left[A_i\right]_{i=1}^j$	-	block diagonal matrix with blocks A_i.

- special vectors and matrices:

$\mathfrak{e}$	=	$[1,\ldots,1]^T$,
T_2	=	$\frac{1}{2}\cdot\mathrm{tridiag}[2\mathfrak{e},-\mathfrak{e}]$,
D_4	=	$4\cdot\mathrm{diag}[\mathfrak{b}]$, where $\mathfrak{b} = \left[i^2+\frac{1}{6}\right]_{i=1}^n$,
C_4	=	$D_4\otimes T_2 + T_2\otimes D_4$,
K_k	-	$\frac{1}{2n^2}C_4$, stiffness matrix for $-x^2u_{yy} - y^2u_{xx}$ using linear finite elements,
$\tilde{C}_{k,r,\mu}$	-	AMLI preconditioner with the polynomial $(1-rt)^\mu$ on level $l = 1,\ldots,k$,
$\bar{C}_{k,S,\mu}$=$\bar{C}_{k,S,\mu,1}$	-	Multi-grid preconditioner (1 iteration) on level k with the smoother S and μ cycles on each level,
$\hat{C}_k$	-	MTS-BPX preconditioner,
$\hat{\mathcal{C}}_k$	-	ILU-BPX preconditioner.

Contents

1 Introduction

Many problems in mechanics, natural sciences, and economy can be described by partial differential equations (pde). Examples are the heat equation of thermodynamics

$$u_t = \triangle u + f,$$

the system of Lamé equations for $\underline{u} = (u^{(1)}, u^{(2)}, u^{(3)})^T$ of linear elasticity

$$-\mu\triangle\underline{u} - (\lambda+\mu)\text{grad div } \underline{u} \;\; = \;\; \underline{f},$$

the Schrödinger equation of quantum mechanics

$$i\hbar\Psi_t = -\frac{\hbar^2}{2m}\triangle\Psi,$$

or the Black-Scholes partial differential equation of pricing of options

$$\frac{\partial v}{\partial t} + \frac{1}{2}\sum_{i,j=1}^{d}\sigma_i\sigma_j\rho_{ij}S_iS_j\frac{\partial^2 v}{\partial S_i\partial S_j} + r\sum_{i=1}^{d}S_i\frac{\partial v}{\partial S_i} - rv \;\; = \;\; 0. \tag{1.1}$$

However, for all these pde's, the exact solution is only known for some academic examples by giving suitably chosen right-hand sides, initial values and boundary values. For the corresponding applications, it is important to obtain solutions of the pde also for those cases in which an exact solution is not known.
For thirty years, applied mathematicians have studied discretization methods to obtain approximate solutions of such pde's. Examples for such approximation methods are the finite difference method (fdm), [69], [41], and the finite element method (fem), [24], [74], [66], [16], which has its origin in the simulation of aerodynamics for aero-planes.
In order to understand the approximation theory, the Poisson equation

$$-\triangle u = f$$

is often used as a reference example. In some cases, the theory can be extended to other examples. E.g. by using Korn's inequality, we obtain the same results for the system of the Lamé equations. For all methods, the described discretizations lead to a system of linear algebraic equations

$$\mathcal{A}\underline{u} = \underline{f}.$$

Using the vector $\underline{u}$, an approximation u_h of the exact solution u can be constructed by the usual finite element isomorphism. The error $u_h - u$ tends to zero in a suitably chosen norm, if the

discretization parameter h tends to zero. Therefore for the practical implementation of such algorithms, it is important to choose a discretization parameter h as small as possible in order to obtain a sufficiently accurate approximation u_h to the exact solution u. Then, the dimension n of the vector $\underline{u} \in \mathbb{R}^n$ will be nearly proportional to h^{-d}, where d is the dimension of the domain in which the partial differential equation is solved. Using finite differences or finite elements of low order (h-version of the fem), the corresponding system matrix $\mathcal{A}$ is a sparse matrix and is positive definite for elliptic problems. More precisely, the number of nonzero elements is of order n. For $d > 1$, the matrix $\mathcal{A}$ has a banded-like structure. For todays computers, it is no problem to store such a sparse matrix of dimensions up to some millions.
Instead of finite difference methods or the h-version of the finite element method, collocation methods, [52], and finite elements of high order (p-version), [72], have become more popular for twenty years. For the h-version of the fem, the polynomial degree p of the shape functions on the elements is kept constant and the mesh-size h is decreased. This is in contrast to the the p-version of the fem in which the polynomial degree p is increased and the mesh-size h is kept constant. The advantage of the p-version in comparison to the h-version is that the approximate solution u_p converges faster to the exact solution u, if u is sufficiently smooth. For example, for the potential equation $-\triangle u = f$ with $u \in C^\infty$, the error in the H^1-Sobolev norm fulfills $\| u - u_p \|_1 \leq Ce^{-rp}$ (with some constant $r > 0$ independent of p) in contrast to the algebraic convergence order of the h-version with $\| u - u_h \|_1 \leq Ch$. Thus, the dimension of the fem ansatz space can be reduced while obtaining an approximate solution with the same accuracy as in the h-version of the fem. Both ideas, mesh refinement and increasing the polynomial degree, can be combined. This is called the hp-version of the fem. Such discretizations lead to a system of algebraic equations

$$\mathcal{A}_p \underline{u}_p = \underline{f}_p \tag{1.2}$$

with $\mathcal{A}_p \in \mathbb{R}^{n_p \times n_p}$, where $n_p \approx p^d$ is the number of unknowns. The question of the solution of such a system $\mathcal{A}_p \underline{u}_p = \underline{f}_p$ is more difficult. The structure of the matrix $\mathcal{A}_p$ depends on the choice of the basis of the fem ansatz space. For the h-version of the fem, it is natural to use Lagrange-interpolation polynomials as basis. The case of the p-version is more delicate. For some kinds of elements, e.g. parallelepipeds in two dimensions, hierarchical polynomials are known for which the matrix $\mathcal{A}_p$ has a sparse structure with $\mathcal{O}(n_p)$ nonzero elements. One example is the basis of the integrated Legendre polynomials.
However, for each parallelepipedian element, the element stiffness matrix $\mathcal{A}_p$ has a banded structure. Therefore, by using direct solvers for (1.2), the memory requirement and the arithmetical cost are not optimal because of fill-in. Hence, iterative methods for (1.2) are better, if p is sufficiently large. In all cases, the matrix $\mathcal{A}_p$ is ill-conditioned which means that the ratio $\frac{\lambda_{max}(\mathcal{A}_p)}{\lambda_{min}(\mathcal{A}_p)}$ is (depending on the choice of the basis) of order $p^2 \ldots p^4$ for $d = 2$ and p^4 or worse for $d = 3$, see e.g. [8], [58]. Thus, an efficient preconditioner for the matrix $\mathcal{A}_p$ is necessary.
Several preconditioners for the p-version of the fem and for the p-version of the boundary element method (bem) have been derived in the last years. Most of them, see [7], [35], [36], [59], [63], [1], [49] for the fem, and [2], [76], [37], [45], [46] for the bem are based on domain decomposition techniques. Efficient solvers for the subproblems are necessary for such a domain decomposition preconditioner. One subproblem solver is the solver for the unknowns corresponding to the

element interfaces which was investigated in two dimensions by Jensen and Korneev, [49], and Ainsworth, [1], and in two and three dimensions by Guo and Cao, [23]. Another ingredient of the domain decomposition preconditioner is the solver related to the interiors of the sub-domains. On the one hand, it is known from the spectral method that the corresponding matrices are spectrally equivalent to matrices resulting from the discretization of the Laplacian using the Gauß-Lobatto points as grid points, [30]. On the other hand, using the basis of scaled integrated Legendre polynomials, this matrix is very similar to discretization matrices of the degenerated elliptic operator $-x^2u_{yy} - y^2u_{xx}$ on the domain $(0,1)^2$, see [9], [53]. Linear or bilinear finite elements on uniform meshes or finite differences on uniform grids are used as discretization method.
For systems of linear algebraic equations resulting from the h-version of the fem, additive and multiplicative solution techniques are known. Examples are multi-grid methods, [40], [43], the BPX preconditioner, [21], [81], domain decomposition methods, [17], [18], [19], [20], and in 2D, the hierarchical basis preconditioner, [80]. In the most convergence proofs for these methods, uniform ellipticity of the differential operator is assumed which is, e.g., for the Laplacian fulfilled. For degenerated operators of the type $-(b(x,y)u_x)_x - u_{yy}$, where $0 < b(x,y) < b_{max}$, Bramble and Zhang proved in [22] a mesh-size independent multi-grid convergence rate $\rho < 1$. However, the operator $-x^2u_{yy} - y^2u_{xx}$ does not satisfy the assumptions of Bramble and Zhang. On the one hand, numerical experiments, see [14] and [11], for discretizations of differential operators as $-x^2u_{yy}-y^2u_{xx}$ indicate a mesh-size independent convergence rate $\rho < 1$ for multi-grid algorithms with semi-coarsening and line-smoother. On the other hand, Braess, [15], Schieweck, [70], and Pflaum, [65], derived purely algebraic techniques in order to prove a mesh-size independent multi-grid convergence rate. There one only has to verify algebraic assumptions which are expressed as eigenvalue estimates of small matrices.
In this work, we will derive arithmetically optimal solvers for several discretization methods of $-x^2u_{yy} - y^2u_{xx} = g$ in the unit square $(0,1)^2$. Moreover, nearly arithmetically optimal preconditioners for the interior problem of the p-version of the fem will be obtained.
The presented work is organized as follows. In chapter 2, some preliminary tools, the theory of simple iterative methods for the solution of (1.2), properties of the integrated Legendre polynomials, and properties of the Kronecker product, will be given. In chapter 3, the discretization of the potential equation by the p-version of the fem will be described. Furthermore, global solution ideas will be derived for the system (1.2). Focusing on the interior problem, which has to be solved by applying domain decomposition preconditioners, several properties of the element stiffness matrix related to a Dirichlet problem will be formulated. Moreover, a first preconditioner for the element stiffness matrix will be proposed. In chapter 4, degenerated elliptic problems in one, two and three dimensions will be investigated. It will be proved that the resulting discretization matrices are equal to the preconditioners of the element stiffness matrix of the p-version of the fem as defined in chapter 3.
In chapter 5, fast multi-level solvers for discretizations of $-x^2u_{yy} - y^2u_{xx}$ in the unit square $(0,1)^2$ will be derived. For this purpose, we will use a sequence of finite element discretizations with piecewise linear shape functions on uniform meshes T_l. The corresponding finite element spaces denoted by $\mathbb{V}_l$ will be split into the direct sum $\mathbb{V}_l = \mathbb{V}_{l-1} \bigoplus \mathbb{W}_l$, $l \geq 2$. A sequence of systems $K_l\underline{u}_l = \underline{g}_l$, $l = 1, \ldots, k$, arises as result of this discretization. In section 5.3, a multi-grid (k-grid) algorithm will be formulated which can be interpreted as alternate, approximate projec-

tion onto the subspaces $\mathbb{V}_{l-1}$ and $\mathbb{W}_l$. Therefore, systems with the matrix K_{l-1} and a matrix $K_{\mathbb{W}_l}$ have to be solved approximately. The matrix $K_{\mathbb{W}_l}$ is the stiffness matrix with respect to the new nodes on level l. The convergence rate σ_k of the considered multi-grid algorithm will be estimated purely algebraic. It depends on the constant in the strengthened Cauchy-inequality and the convergence rate of the iterative procedure in order to solve $K_{\mathbb{W}_l}\underline{w} = \underline{r}$. For the system $K_{\mathbb{W}_l}\underline{w} = \underline{r}$, a special line smoother $S_{0,l}$ will be defined whose error transion operator is given by $I - C_{\mathbb{W}_l}^{-1}K_{\mathbb{W}_l}$. Finally, it will be proved that the convergence rate σ_k of the multi-grid algorithm for the solution of $K_k\underline{w} = \underline{r}$ is bounded by a constant $\sigma < 1$. Using the matrices $C_{\mathbb{W}_l}$, an Algebraic Multi-level Iteration (AMLI) preconditioner, [5], [6], will be proposed. In section 5.4, condition number estimates for the AMLI preconditioned systems will be given. Section 5.5 will investigate multi-grid algorithms using different smoothers for $-x^2u_{yy} - y^2u_{xx}$ and similar problems. A smoother similar to the smoother $S_{0,l}$ will be introduced. In numerical experiments, the application of this smoother instead of $S_{0,k}$ embedded in the multi-grid algorithm will accelerate the convergence of the algorithm. However, a convergence result cannot be proved. Moreover, a symmetric and positive definite multi-grid preconditioner will be derived. It will be shown that the condition number of the preconditioned system is bounded by a constant independent of the mesh-size h. In section 5.6, a BPX-like preconditioner, which we call MTS-BPX preconditioner, will be introduced. This preconditioner can be interpreted as BPX preconditioner with smoothing. It will be proved that the upper eigenvalue of the MTS-BPX preconditioned system is bounded by ck (k level number) in the case of piecewise linear fem-discretizations for differential operators of the type $-x^\alpha u_{yy} - y^\alpha u_{xx}$ ($\alpha \geq 0$). In section 5.7, it will be proved that one iteration of all proposed algorithms is arithmetically optimal. Moreover, an interpretation of the smoother $S_{0,k}$ as line-smoother will be shown. Finally, numerical experiments of all methods will be given in section 5.8.
In chapter 6, the preconditioners for the element stiffness matrix of the p-version of the fem in two dimensions will be defined. The main condition number estimates will be given. Furthermore, all proposed preconditioners will be compared numerically. In chapter 7, some new ideas concerning preconditioning the element stiffness matrix of the p-version of the fem in two and three dimensions using wavelet bases will be formulated.

2 Preliminary Tools

2.1 Iterative solution methods for systems of linear equations

The aim of this section is to consider iterative methods in order to solve a system of linear algebraic equations. Furthermore, the convergence properties of several iterative methods are shown and the purpose of an effective preconditioning is motivated.

2.1.1 Simple iterative methods

Most simple iterative methods, [40], [3], [62], in order to solve a system of linear equations

$$\mathcal{A}\underline{x} = \underline{b} \tag{2.1.1}$$

can be written as Richardson-iteration, i.e. the new iterate $\underline{x}^{(m+1)}$ is given by the recursion

$$\underline{x}^{(m+1)} = \underline{x}^{(m)} - \omega\mathcal{C}^{-1}(\mathcal{A}\underline{x}^{(m)} - \underline{b}). \tag{2.1.2}$$

The parameter ω is a damping parameter and the matrix $\mathcal{C}$ is a good approximation to the matrix $\mathcal{A}$. The matrix $\mathcal{C}$ is called a preconditioning matrix. Choosing $\mathcal{C} = \mathcal{D}$, where $\mathcal{D}$ is the diagonal part of $\mathcal{A}$, one obtains the ω-Jacobi method, whereas the ω-Gauß-Seidel, or SOR, method is defined with the choice of $\mathcal{C} = \mathcal{D} + \omega\mathcal{L}$. The matrix $\mathcal{L}$ is the strongly lower triangular part of $\mathcal{A}$. The speed of convergence of the sequence $\{\underline{x}^{(m)}\}_{m=1}^{\infty}$ to the exact solution $\underline{x}^*$ of (2.1.1) depends on the condition number of the matrix $\mathcal{C}^{-1}\mathcal{A}$: One obtains

$$\underline{x}^{(m+1)} - \underline{x}^* = (I - \omega\mathcal{C}^{-1}\mathcal{A})(\underline{x}^{(m)} - \underline{x}^*)$$

by adding $-\underline{x}^*$ to (2.1.2) and $\mathcal{A}\underline{x}^* = \underline{b}$. Therefore, the convergence rate in the Euclidian norm is given gy

$$\sup_{\underline{x}^{(m)} \neq \mathbf{0}} \frac{\| \underline{x}^{(m+1)} - \underline{x}^* \|_2}{\| \underline{x}^{(m)} - \underline{x}^* \|_2} = \rho(I - \omega\mathcal{C}^{-1}\mathcal{A}),$$

where the parameter $\rho(\mathcal{B})$ is the spectral radius of the matrix $\mathcal{B} = I - \omega\mathcal{C}^{-1}\mathcal{A}$. If $\rho(I - \omega\mathcal{C}^{-1}\mathcal{A}) \geq 1$, the method does not converge to the exact solution $\underline{x}^*$. Let us assume that $\mathcal{A}$ and $\mathcal{C}$ are symmetric and positive definite. Thus, $\mathcal{C}^{-1}\mathcal{A}$ has positive real eigenvalues and the optimal damping parameter $\omega = \omega_{opt}$ is given by

$$\omega = \omega_{opt} = \frac{2}{\lambda_{max}(\mathcal{C}^{-1}\mathcal{A}) + \lambda_{min}(\mathcal{C}^{-1}\mathcal{A})},$$

see e.g. [40], [3], [62]. Hence, inserting this value for ω, one achieves

$$\rho(I - \omega\mathcal{C}^{-1}\mathcal{A}) = \frac{\lambda_{max}(\mathcal{C}^{-1}\mathcal{A}) - \lambda_{min}(\mathcal{C}^{-1}\mathcal{A})}{\lambda_{max}(\mathcal{C}^{-1}\mathcal{A}) + \lambda_{min}(\mathcal{C}^{-1}\mathcal{A})}. \tag{2.1.3}$$

Note that for a symmetric and positive definite matrix $\mathcal{B} \in \mathbb{R}^{n\times n}$,

$$\| \mathcal{B} \|_2 = \sqrt{\rho(\mathcal{B}^*\mathcal{B})} = \lambda_{max}(\mathcal{B}).$$

Thus, $\kappa(\mathcal{B}) = \| \mathcal{B} \|_2 \| \mathcal{B}^{-1} \|_2 = \lambda_{max}(\mathcal{B})/\lambda_{min}(\mathcal{B})$. Moreover, the symmetric and positive definite matrix $\mathcal{B} = \mathcal{C}^{-1/2}\mathcal{A}\mathcal{C}^{-1/2}$ has the same eigenvalues as $\mathcal{C}^{-1}\mathcal{A}$. Hence by definition, let $\kappa\left(\mathcal{C}^{-1}\mathcal{A}\right) = \frac{\lambda_{max}\left(\mathcal{C}^{-1}\mathcal{A}\right)}{\lambda_{min}\left(\mathcal{C}^{-1}\mathcal{A}\right)}$ be the condition number of the matrix $\mathcal{C}^{-1/2}\mathcal{A}\mathcal{C}^{-1/2}$. Therefore, by relation (2.1.3),

$$\rho(I - \omega\mathcal{C}^{-1}\mathcal{A}) = \frac{\kappa(\mathcal{C}^{-1}\mathcal{A}) - 1}{\kappa(\mathcal{C}^{-1}\mathcal{A}) + 1},$$

which means that the condition number κ of $\mathcal{C}^{-1}\mathcal{A}$ should be small in order to achieve a small $\rho(I - \omega\mathcal{C}^{-1}\mathcal{A}) < 1$. Let m be the number of iterations in order to obtain a relative accuracy of ε in the Euclidian norm, i.e. m is the smallest integer with $\| \underline{x}^{(m)} - \underline{x}^* \|_2 \leq \varepsilon \| \underline{x}^{(0)} - \underline{x}^* \|_2$. Then,

$$m \leq 2\kappa\left(\mathcal{C}^{-1}\mathcal{A}\right) |\log \varepsilon|.$$

Hence, for the convergence of an iterative method of the type (2.1.2), a matrix $\mathcal{C}$ has to be constructed which satisfies the following two conditions:

- the condition number of $\mathcal{C}^{-1}\mathcal{A}$ should be small,
- due to (2.1.2) for each iteration step, the operation $\underline{w} = \mathcal{C}^{-1}\underline{r}$ should be cheap.

In general, this problem cannot be solved satisfactory. However, nowadays there are several ideas deriving preconditioners $\mathcal{C}$ using the origin of the matrix $\mathcal{A}$.

2.1.2 Pcg-method

The preconditioned conjugate gradient method for the solution of $\mathcal{A}\underline{x} = \underline{b}$ with symmetric and positive definite $\mathcal{A}$ has been developed by Hestenes and Stiefel, [44]. It is a Krylov subspace method. It can be used as a direct method because it gives theoretically the exact solution after n iterations, where n is the dimension of the system. Because of its fast convergence properties, it is used as an iterative method. Let $\mathcal{C}$ be a symmetric and positive definite matrix (preconditioner for $\mathcal{A}$). The sequence $\{\underline{x}^{(m)}\}_{m=1}^{\infty}$ will be computed as follows, see e.g. [40], [3], [62], [73],

– Initialization:

- $\underline{r}^{(0)} = \mathcal{A}\underline{x}^{(0)} - \underline{b}$,
- $\underline{w}^{(0)} = \underline{q}^{(1)} = \mathcal{C}^{-1}\underline{r}^{(0)}$,

- $\gamma_0 = \left(\underline{w}^{(0)}, \underline{r}^{(0)}\right)$.

– Iteration: For $m = 1, \ldots,$ do

- $\underline{v}^{(m)} = \mathcal{A}\underline{w}^{(m-1)}$,
- $\delta_m = \left(\underline{v}^{(m)}, \underline{q}^{(m-1)}\right)$, $\alpha_m = \frac{\gamma_{m-1}}{\delta_m}$,
- $\underline{x}^{(m)} = \underline{x}^{(m-1)} + \alpha_m \underline{q}^{(m-1)}$,
- $\underline{r}^{(m)} = \underline{r}^{(m-1)} + \alpha_m \underline{v}^{(m)}$,
- $\underline{w}^{(m)} = \mathcal{C}^{-1}\underline{r}^{(m)}$,
- $\gamma_m = \left(\underline{w}^{(m)}, \underline{r}^{(m)}\right)$, $\beta_m = \frac{\gamma_{m-1}}{\gamma_m}$,
- $\underline{q}^{(m)} = \underline{w}^{(m)} + \beta_m \underline{q}^{(m-1)}$.

Let $\underline{x}^*$ be the exact solution of (2.1.1). Then, the following convergence result can be shown for the sequence $\{\underline{x}^{(m)}\}_{m=1}^{\infty}$. Let

$$\rho = \sup_{\underline{x}^{(m)} \neq \mathbf{0}} \frac{\| \underline{x}^{(m+1)} - \underline{x}^* \|_{\mathcal{A}}}{\| \underline{x}^{(m)} - \underline{x}^* \|_{\mathcal{A}}}.$$

Then, see [40], [3], [62], [73], the relation

$$\rho \leq 2\frac{\sqrt{\kappa\left(\mathcal{C}^{-1}\mathcal{A}\right)} - 1}{\sqrt{\kappa\left(\mathcal{C}^{-1}\mathcal{A}\right)} + 1}$$

is valid. This means that the number m of iterations in order to achieve a relative accuracy of ε is bounded by

$$m \leq \frac{1}{2}\sqrt{\kappa\left(\mathcal{C}^{-1}\mathcal{A}\right)} \ln\left(\frac{2}{\varepsilon}\right) + 1.$$

Hence, the numbers m of iterations grow proportionally to $\sqrt{\kappa\left(\mathcal{C}^{-1}\mathcal{A}\right)}$ in contrast to $\kappa\left(\mathcal{C}^{-1}\mathcal{A}\right)$ for the most simple iterative methods.

2.2 Cholesky decomposition for banded matrices and related methods

In this section, the memory requirement M_n and the number of operations $\mathfrak{W}_n$ in order to solve the linear system

$$\mathcal{A}\underline{x} = \underline{b}$$

with $\mathcal{A} \in \mathbb{R}^{n \times n}$ symmetric and positive definite are considered. We assume that $\mathcal{A} = [a_{ij}]_{i,j=1}^n$ has a banded structure with bandwidth m, i.e. $a_{ij} = 0$ for $|i-j| > m$. Determining the Cholesky decomposition $\mathcal{A} = \mathcal{L}\mathcal{L}^T$, [73], [34], with the lower triangular matrix $\mathcal{L} = [l_{ij}]_{i,j=1}^n$, one obtains the relations $l_{ij} = 0$ for $|i-j| > m$. However, the relation $l_{ij} = 0$ is not necessarily satisfied, if

$a_{ij} = 0$ and $|i - j| < m$. Therefore, one obtains $M_n \preceq nm$ and the cost for the computation of $\mathcal{L}$ is $\mathfrak{W}_n \preceq nm^2$.
A special case is that $\mathcal{A}$ is a symmetric and positive definite tridiagonal matrix. Note, that $m = 1$ holds. Then $M_n \asymp n$ and $\mathfrak{W}_n \asymp n$, i.e. the Cholesky decomposition is arithmetically optimal. However for matrices of five-point stencil structure, the relation $a_{ij} = 0$ holds, if $|i - j| \notin \{0, 1, m\}$, where $m^2 = n$. One obtains only $\mathfrak{W}_n \asymp nm^2 = n^2$ and $M_n \asymp n^{\frac{3}{2}}$, which means that the memory requirement is $\mathcal{O}(n^{\frac{3}{2}})$ in order to save $\mathcal{L}$, whereas the memory requirement is about $5n$ in order to save $\mathcal{A}$. Thus, the Cholesky decomposition is not optimal.
In the seventies, several other direct methods are derived for five- and nine- point stencils. These methods reorder the unknowns in such a way that the Cholesky decomposition of the reordered matrix produces less fill-in than the Cholesky decomposition of the usual matrix. The asymptotically most efficient one is the method of Nested Dissection developed by George [33], [32]. In this method, the corresponding graph (G, V) of the matrix is considered. A separator S of the graph G is constructed which divides the graph into the disjoint subgraphs (G_1, V_1) and (G_2, V_2) with $G = V_1 \cup V_2 \cup S$ and $V_i \cap S = \emptyset$, $i = 1, 2$. Now, the vertices of V_1 are ordered firstly, next the vertices of V_2 and as last those of S. Doing this algorithm recursively for G_1 and G_2, a new ordering of the vertices is given. Then, the arithmetical cost can be reduced to $\mathcal{O}(m^3)$, the memory requirement to $\mathcal{O}(m^2 \log(1+m))$, [33]. However, this is not optimal, i.e. $\mathfrak{W}_m > \mathcal{O}(m^2)$ and $M_m > \mathcal{O}(m^2)$. The integer m denotes the number of grid-points in one direction.

2.3 Properties of the Legendre polynomials

In this section, the Legendre polynomials are introduced and their most important orthogonality relations are given. We refer to [77] for more facts. Let

$$L_i(x) = \frac{1}{2^i i!} \frac{\mathrm{d}^i}{\mathrm{d}x^i}(x^2 - 1)^i \tag{2.3.1}$$

be the i-th Legendre polynomial and

$$\hat{L}_i(x) = \gamma_i \int_{-1}^{x} L_{i-1}(s)\, \mathrm{d}s, \quad i \geq 2 \tag{2.3.2}$$

be the i-th integrated Legendre polynomial with the scaling factor

$$\gamma_i = \sqrt{\frac{(2i-3)(2i-1)(2i+1)}{4}}. \tag{2.3.3}$$

Moreover, let by definition

$$\begin{aligned} \hat{L}_0(x) &= \frac{1-x}{2}, \\ \hat{L}_1(x) &= \frac{1+x}{2}. \end{aligned}$$

LEMMA 2.1. *The following relations are valid between the polynomials (2.3.1) and (2.3.2):*

$$\frac{\mathrm{d}}{\mathrm{d}x}\hat{L}_i(x) = \gamma_i L_{i-1}(x), \quad i \geq 2, \tag{2.3.4}$$

$$\int_{-1}^{1} L_i(x)L_j(x)\,\mathrm{d}x = \delta_{ij}\frac{2}{2i+1}, \quad i \geq 0, \tag{2.3.5}$$

$$(2i+1)L_i(x) = \frac{\mathrm{d}}{\mathrm{d}x}(L_{i+1}(x) - L_{i-1}(x)), \quad i \geq 1, \tag{2.3.6}$$

$$L_i(-1) = (-1)^i, \quad i \geq 0, \tag{2.3.7}$$

$$\hat{L}_i(x) = \sqrt{\frac{(2i+1)(2i-3)}{4(2i-1)}}(L_i(x) - L_{i-2}(x)), \quad i \geq 2, \tag{2.3.8}$$

$$\hat{L}_i(1) = 0, \quad i \geq 2, \tag{2.3.9}$$

$$\hat{L}_i(-1) = 0, \quad i \geq 2, \tag{2.3.10}$$

$$(i+1)L_{i+1}(x) + iL_{i-1}(x) = (2i+1)xL_i(x), \quad i \geq 1. \tag{2.3.11}$$

Proof: The proof is given in [77].

2.4 Kronecker product

In this work, several properties of the Kronecker product are used. The most important are summarized in this section.

DEFINITION 2.2. *Let $A \in \mathbb{C}^{k\times l}$ and $B \in \mathbb{C}^{m\times n}$. Then, the matrix*

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1l}B \\ a_{21}B & a_{22}B & \cdots & a_{2l}B \\ \vdots & & \ddots & \\ a_{k1}B & a_{k2}B & \cdots & a_{kl}B \end{pmatrix} \in \mathbb{C}^{km\times ln} \tag{2.4.1}$$

is called the Kronecker-product between the matrices A and B.

LEMMA 2.3. *Let $A \in \mathbb{C}^{k\times l}$ and $B \in \mathbb{C}^{m\times n}$. Furthermore, let $\alpha \in \mathbb{C}$ and $C \in \mathbb{C}^{k\times l}$, $D \in \mathbb{C}^{l\times s}$ and $E \in \mathbb{C}^{n\times r}$. The following relations are valid:*

$$(\alpha A) \otimes B = A \otimes (\alpha B) = \alpha(A \otimes B), \tag{2.4.2}$$

$$(A \otimes B)^T = A^T \otimes B^T, \tag{2.4.3}$$

$$(A + C) \otimes B = A \otimes B + C \otimes B, \tag{2.4.4}$$

$$(A \otimes B)(D \otimes E) = (AD) \otimes (BE), \tag{2.4.5}$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}, \tag{2.4.6}$$

where the matrices $A \in \mathbb{C}^{l\times l}$ and $B \in \mathbb{C}^{n\times n}$ are non-singular in (2.4.6).

Proof: The proof can be found in several books about Linear Algebra, see, e.g. [64].□
In the following, we assume that $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$. The eigenvalues and eigenvectors of $A \otimes B$ can be determined by the eigenvalues of A and B.

LEMMA 2.4. *Let λ_i^A be an eigenvalue and $\underline{x}_i^A$ the corresponding eigenvector of A, λ_j^B and $\underline{x}_j^B$ an eigenpair of B. Then, $\lambda_i^A\lambda_j^B$ is an eigenvalue of $A \otimes B$ with the eigenvector $\underline{x}_i^A \otimes \underline{x}_j^B$.*

Proof: The assertion follows from (2.4.2).□
The next lemma will be used very often in this work.

LEMMA 2.5. *Let $n_1 = n_3$ and $n_2 = n_4$. Let us assume that the matrices $A_i \in \mathbb{R}^{n_i\times n_i}$, and $B_i \in \mathbb{R}^{n_i\times n_i}$, $i = 1, 2, 3, 4$, are symmetric and positive definite. Furthermore, let*

$$\lambda_{min}\left({B_i}^{-1}A_i\right) \geq \lambda_i, \qquad \lambda_{max}\left({B_i}^{-1}A_i\right) \leq \lambda^i$$

for $i = 1, 2, 3, 4$. Moreover, let

$$\begin{aligned} A &= \alpha A_1 \otimes A_2 + \beta A_3 \otimes A_4, \\ B &= \alpha B_1 \otimes B_2 + \beta B_3 \otimes B_4, \end{aligned}$$

where $\alpha, \beta > 0$. Then, the following eigenvalue estimates are valid

$$\begin{aligned} \lambda_{min}\left(B^{-1}A\right) &\geq \min\left\{\lambda_1\lambda_2, \lambda_3\lambda_4\right\}, \\ \lambda_{max}\left(B^{-1}A\right) &\leq \max\left\{\lambda^1\lambda^2, \lambda^3\lambda^4\right\}. \end{aligned}$$

Proof: Note that by Lemma 2.3,

$$(B_i \otimes B_j)^{-1}(A_i \otimes A_j) = B_i^{-1}A_i \otimes B_j^{-1}A_j.$$

Thus, by Lemma 2.4,

$$\lambda_i\lambda_j \leq \lambda_{min}\left((B_i \otimes B_j)^{-1}(A_i \otimes A_j)\right)$$

and

$$\lambda_{max}\left((B_i \otimes B_j)^{-1}(A_i \otimes A_j)\right) \leq \lambda^i\lambda^j.$$

By our assumptions, the matrices A_i and B_i, $i = 1, 2, 3, 4$, are symmetric and positive definite. Thus, one concludes

$$\lambda_1\lambda_2\left(B_1 \otimes B_2\underline{v}, \underline{v}\right) \leq \left(A_1 \otimes A_2\underline{v}, \underline{v}\right) \leq \lambda^1\lambda^2\left(B_1 \otimes B_2\underline{v}, \underline{v}\right) \tag{2.4.7}$$

and

$$\lambda_3\lambda_4\left(B_3 \otimes B_4\underline{v}, \underline{v}\right) \leq \left(A_3 \otimes A_4\underline{v}, \underline{v}\right) \leq \lambda^3\lambda^4\left(B_3 \otimes B_4\underline{v}, \underline{v}\right) \tag{2.4.8}$$

for all $\underline{v} \in \mathbb{R}^{n_1n_2}$. Multiplying (2.4.7) by $\alpha > 0$, (2.4.8) by $\beta > 0$ and adding both inequalities gives

$$\begin{aligned} \min\left\{\lambda_1\lambda_2, \lambda_3\lambda_4\right\}\left((\alpha B_1 \otimes B_2 + \beta B_3 \otimes B_4)\underline{v}, \underline{v}\right) &\leq \\ \alpha\lambda_1\lambda_2\left(B_1 \otimes B_2\underline{v}, \underline{v}\right) + \beta\lambda_3\lambda_4\left(B_3 \otimes B_4\underline{v}, \underline{v}\right) &\leq \left(A\underline{v}, \underline{v}\right) \end{aligned}$$

and

$$\begin{aligned} \left(A\underline{v}, \underline{v}\right) &\leq \alpha\lambda_1\lambda_2\left(B_1 \otimes B_2\underline{v}, \underline{v}\right) + \beta\lambda_3\lambda_4\left(B_3 \otimes B_4\underline{v}, \underline{v}\right) \\ &\leq \max\left\{\lambda_1\lambda_2, \lambda_3\lambda_4\right\}\left((\alpha B_1 \otimes B_2 + \beta B_3 \otimes B_4)\underline{v}, \underline{v}\right) \end{aligned}$$

for all $\underline{v} \in \mathbb{R}^{n_1n_2}$ which is the desired result. □

3 Discretization by the p-version of the fem

In this chapter, the discretization of the potential equation in two dimensions by the p-version of the fem is investigated. In the next sections, the derivation of the system of linear algebraic equations and first general ideas, namely domain decomposition techniques, in order to solve such a system are explained. One ingredient of such a domain decomposition preconditioner, the solver for the interior problem, will be focused in sections 3.3 and 3.4.

3.1 Formulation of the problem in two dimensions

We consider the boundary value problem

$$\begin{aligned} -\triangle u &= f \quad \text{in} \quad \Omega_1, \\ u &= 0 \quad \text{on} \quad \Gamma_1, \\ \frac{\partial u}{\partial n} &= 0 \quad \text{on} \quad \Gamma_2, \end{aligned} \tag{3.1.1}$$

where $\Omega_1 \subset \mathbb{R}^2$ is a domain which can be decomposed into (straight-line) quadrilaterals and $\Gamma_1 \cup \Gamma_2 = \partial\Omega_1$, $\Gamma_1 \cap \Gamma_2 = \emptyset$. The weak formulation of this problem is:
Find $u \in H_0(\Omega_1) := \{u \in H^1(\Omega_1), u\,|_{\Gamma_1} = 0\}$ such that

$$a_\triangle(u,v) := \int_{\Omega_1} u_x v_x + u_y v_y = \int_{\Omega_1} fv \quad \forall v \in H_0(\Omega_1) \tag{3.1.2}$$

holds. Problem (3.1.1) will be discretized by means of the p-version of the finite element method using quadrilaterals R_s. Let $\mathcal{R}_2 = (-1,1)^2$ be the reference element and $\Phi_s : \mathcal{R}_2 \to R_s$ be the bilinear mapping to the element R_s. We define the finite element space

$$\mathbb{M} := \{u \in H_0(\Omega_1), u\,|_{R_s} = u(\Phi_s(\xi,\eta)) = \tilde{u}(\xi,\eta), \tilde{u} \in \mathbb{Q}_p\},$$

where $\mathbb{Q}_p$ is the space of all polynomials $p(\xi,\eta) = p_1(\xi)p_2(\eta)$ of maximal degree p in each variable. Now, the discretized problem can be formulated: Find $u^p \in \mathbb{M}$ such that

$$a_\triangle(u^p, v^p) = \int_{\Omega_1} fv^p \quad \forall v^p \in \mathbb{M} \tag{3.1.3}$$

holds. Let $(\psi_1, \ldots, \psi_{n_p})$ be a basis of $\mathbb{M}$. Then, problem (3.1.3) is equivalent to solving the system of algebraic finite element equations

$$A_p \underline{u}_p = \underline{f}_p, \tag{3.1.4}$$

where

$$\begin{aligned} A_p &= [a_\triangle(\psi_j, \psi_i)]_{i,j=1}^{n_p}, \\ \underline{u}_p &= [u_i]_{i=1}^{n_p}, \\ \underline{f}_p &= \left[\int_{\Omega_1} f\psi_i\right]_{i=1}^{n_p}. \end{aligned}$$

Then, $u^p = \sum_i u_i\psi_i$ is the solution of (3.1.3). We are interested in finding an efficient solver for the system of linear algebraic equations (3.1.4).

3.2 Domain decomposition

Domain decomposition techniques, [63], [17], [18], [19], [20], [60], [61], are efficient iterative methods in order to solve linear systems of algebraic equations of the type (3.1.4). The approximation space $\mathbb{M}$ will be split into a direct sum $\mathbb{M} = \mathbb{M}_1 \oplus \ldots \oplus \mathbb{M}_k$. It is assumed that this splitting is stable with respect to the bilinear form $a_\triangle$, i.e. the relation

$$\sum_{i=1}^{k} a_\triangle(v_i, v_i) \leq c^2 a_\triangle(v, v)$$

is valid for all $v_i \in \mathbb{M}_i$ and $v = \sum_{i=1}^{k} v_i$. The efficient preconditioner

$$\mathcal{C}^{-1} = \sum_{i=1}^{k} V_i (V_i^T A_p V_i)^{-1} V_i^T$$

can be built, where V_i is the matrix representation of the orthogonal projection $\mathbb{M} \mapsto \mathbb{M}_i$ with respect to the energetic scalar product $a_\triangle(\cdot,\cdot)$. Note that $V_i^T A_p V_i$ is the stiffness matrix of:
Find $v_i \in \mathbb{M}_i$ such that

$$a_\triangle(v_i, w_i) = \langle f, w_i \rangle \quad \forall w_i \in \mathbb{M}_i.$$

Then, the eigenvalue estimates $\lambda_{min}\left(\mathcal{C}^{-1}A_p\right) \geq c^{-2}$ and $\lambda_{max}\left(\mathcal{C}^{-1}A_p\right) \leq k$ are valid, cf. [56], [55].
For our purpose, we have to choose $k = 3$. The corresponding spaces are defined as follows:

- $\mathbb{M}_1 = \mathbb{M}_{vert}$ is the space of the vertex functions which are the usual piecewise bilinear functions of the h-version of the finite element method,
- $\mathbb{M}_2 = \mathbb{M}_{edg}$ is the space of the edge bubble functions,
- $\mathbb{M}_3 = \mathbb{M}_{int}$ is the space of the interior bubbles which are nonzero on one element only.

An edge bubble function corresponds to an edge e of the mesh. Its support is formed by those two elements which have this edge e in common. Corresponding to this splitting of the shape functions, the matrix A_p is split analogously into sub-blocks,

$$A_p = \begin{bmatrix} A_{vert} & A_{vert,edg} & A_{vert,int} \\ A_{edg,vert} & A_{edg} & A_{edg,int} \\ A_{int,vert} & A_{int,edg} & A_{int} \end{bmatrix}. \tag{3.2.1}$$

The indices $vert$, edg and int denote the blocks corresponding to the vertex, edge bubble and interior bubble functions, respectively. Jensen and Korneev, [49], and Ivanov and Korneev, [47], [48], developed preconditioners for the p-version of the finite element method in a two-dimensional domain using domain decomposition techniques, [7]. They proposed the preconditioning matrix

$$C_p = \begin{bmatrix} A_{vert} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A_{edg} & A_{edg,int} \\ \mathbf{0} & A_{int,edg} & A_{int} \end{bmatrix} \tag{3.2.2}$$

corresponding to the splitting $\mathbb{M}_{vert} \oplus (\mathbb{M}_{edg} \oplus \mathbb{M}_{int})$ which is considered in a first step. This splitting is nearly stable as the following lemma confirms.

LEMMA 3.1. *The condition number* $\kappa\left({C_p}^{-1} A_p\right)$ *grows as* $(1 + \log p)$.

Proof: The proof can be found in [47], Lemma 2.3. □
Therefore, the vertex unknowns can be determined separately. Efficient solution methods are direct solvers in the case of the p-version of the fem, if the matrix A_{vert} is small, or multi-grid methods, [40], in the hp-version. However, the splitting $\mathbb{M}_{edg} \oplus \mathbb{M}_{int}$ is not stable. Therefore, we can proceed as follows. The sub-block corresponding to $\mathbb{M}_{edg}$ and $\mathbb{M}_{int}$ is factorized as

$$\begin{bmatrix} A_{edg} & A_{edg,int} \\ A_{int,edg} & A_{int} \end{bmatrix} = \begin{bmatrix} I & A_{edg,int}A_{int}^{-1} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \hat{S} & \mathbf{0} \\ \mathbf{0} & A_{int} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ A_{int}^{-1}A_{int,edg} & I \end{bmatrix}$$

with the Schur complement

$$\hat{S} := A_{edg} - A_{edg,int}A_{int}^{-1}A_{int,edg}.$$

Thus for $\mathbb{M}_{int}$, the subproblem restricted to this space has to be solved, whereas for $\mathbb{M}_{edg}$ a modified problem is considered. The matrix A_{int} corresponds to the interior bubbles having a support containing one element only. Therefore, the matrix A_{int} is a block diagonal matrix, where each block corresponds to one element. Hence, in order to compute the interior unknowns, we have to solve a Dirichlet problem on each quadrilateral. The edge unknowns are computed via the Schur complement $\hat{S}$ and multiplications with the matrix $\begin{bmatrix} I \\ -A_{int}^{-1}A_{int,edg} \end{bmatrix}$ and its transpose. So, in addition to a solver for A_{vert}, three tools are required to define a preconditioner for the matrix of (3.2.2), namely

- a preconditioner for the interior problem,

- a preconditioner for the Schur complement $\hat{S}$ and
- an extension operator from the edges of a quadrilateral into its interior in order to replace the matrix $A_{int}^{-1}A_{int,edg}$.

Ivanov and Korneev, [47], [48], derived some preconditioners $C_{\hat{S}}$ for the Schur complement. The condition number of $C_{\hat{S}}^{-1}\hat{S}$ is $\mathcal{O}(1+\log^2 p)$ in the worst case, where p is the polynomial degree. The solution of $C_{\hat{S}}\underline{x} = \underline{y}$ can be done by solving triangular systems and fast Fourier transform, [28]. The problem of the extension operator was investigated by Babuška et. al, [7].
We focus now on a fast solver for $A_{int} = \text{blockdiag}\,[A_{R_s}]_s$, where A_{R_s} is that block of the stiffness matrix A_{int} which corresponds to the element R_s. The following lemma is valid.

LEMMA 3.2. *Let $\partial R_s \in C^{(t)}$, $t \geq 2$, where $C^{(t)}$ denotes the class of all boundaries which consist of a finite number of t times continuously differentiable curves and the angles of these curves at their intersection points on ∂R_s are distinct from 0 and 2π. Then, $\kappa\left({A_{R_s}}^{-1}A_{\mathcal{R}_2}\right) = \mathcal{O}(1)$, where $A_{\mathcal{R}_2} = (-1,1)^2$.*

Proof: The proof can be found in [49], Lemma 4.2. □
Hence, it is sufficient to investigate the matrix $A_{\mathcal{R}_2}$ in order to find a good preconditioner for A_{int}. This will be done in the next sections and chapters.

3.3 Properties of the element stiffness matrix

Let $d = 2$ be the dimension of the domain. By Lemma 3.2,

$$\begin{array}{rcl} -\triangle u & = & f \quad \text{in} \quad \mathcal{R}_d = (-1,1)^d, \\ u & = & 0 \quad \text{on} \quad \partial\mathcal{R}_d \end{array} \tag{3.3.1}$$

is the typical model problem in order to solve the system

$$A_{int}\underline{x} = \underline{y}$$

of linear algebraic finite element equations. Problem (3.3.1) will be investigated in the case $d = 3$ as well. Problem (3.3.1) is solved by the $p-$version of the finite element method with one element $\mathcal{R}_d$ only. As finite element space,

$$\mathbb{M}_p = \begin{cases} H_0^1(\mathcal{R}_2) \cap \text{span}\{\phi_{ij}(x,y)\}_{i,j=0}^{p} & \text{for} \quad d = 2, \\ H_0^1(\mathcal{R}_3) \cap \text{span}\{\phi_{ijk}(x,y,z)\}_{i,j,k=0}^{p} & \text{for} \quad d = 3 \end{cases}$$

is chosen, where $\phi_{ij}(x,y) = x^i y^j$ and $\phi_{ijk}(x,y,z) = x^i y^j z^k$, respectively. The discrete problem is: Find $u^p \in \mathbb{M}_p$ such that

$$\int_{\mathcal{R}_d} \nabla u^p \cdot \nabla v^p = \int_{\mathcal{R}_d} f v^p \quad \forall v^p \in \mathbb{M}_p. \tag{3.3.2}$$

In order to define a basis in $\mathbb{M}_p$, we choose tensor products of the integrated Legendre polynomials $\hat{L}_i$ (2.3.2). More precisely, let

$$\begin{array}{rcll} \hat{L}_{ij}(x,y) & = & \hat{L}_i(x)\hat{L}_j(y) & 0 \le i,j \le p, \\ \hat{L}_{ijk}(x,y,z) & = & \hat{L}_i(x)\hat{L}_j(y)\hat{L}_k(z) & 0 \le i,j,k \le p. \end{array}$$

Since $\hat{L}_i(\pm 1) = 0$ for $i \ge 2$, cf. relations (2.3.9) and (2.3.10),

$$\mathbb{M}_p = \operatorname{span}\{\hat{L}_{ij}(x,y)\}_{i,j=2}^{p}$$

for $d = 2$ and

$$\mathbb{M}_p = \operatorname{span}\{\hat{L}_{ijk}(x,y,z)\}_{i,j,k=2}^{p}$$

for $d = 3$. The stiffness matrix $A_{\mathcal{R}_2}$ for (3.3.2) (with $d = 2$) is given by $A_{\mathcal{R}_2} = [a_{ij,kl}]_{i,j=2;k,l=2}^{p}$, where

$$a_{ij,kl} = \int_{\mathcal{R}_2} \nabla\hat{L}_{ij}(x,y) \cdot \nabla\hat{L}_{kl}(x,y)\ \mathrm{d}(x,y). \tag{3.3.3}$$

Analogously, the matrix $A_{\mathcal{R}_3}$ is defined. The matrices $A_{\mathcal{R}_d}$ can be written explicitely as

$$\begin{array}{rcll} A_{\mathcal{R}_2} & = & F \otimes D + D \otimes F & \text{and} \\ A_{\mathcal{R}_3} & = & F \otimes F \otimes D + F \otimes D \otimes F + D \otimes F \otimes F, & \end{array} \tag{3.3.4}$$

where the matrices F and D are the one-dimensional mass matrix and stiffness matrix in the basis of the integrated Legendre polynomials $\{\hat{L}_i(x)\}_{i=2}^{p}$, i.e.

$$\begin{aligned} F &= \left[\int_{-1}^{1} \hat{L}_i(x)\hat{L}_k(x)\ \mathrm{d}x\right]_{i,k=2}^{p}, \\ D &= \left[\int_{-1}^{1} \hat{L}_i'(x)\hat{L}_k'(x)\ \mathrm{d}x\right]_{i,k=2}^{p}. \end{aligned}$$

Then using relations (2.3.4), (2.3.5) and (2.3.8), a simple calculation shows

$$\begin{array}{rcl} F & = & \operatorname{pentdiag}[\mathfrak{e}, \mathbf{0}, \mathfrak{p}], \\ D & = & \operatorname{diag}[\mathfrak{d}] \end{array} \tag{3.3.5}$$

with the coefficients

$$\begin{aligned} \mathfrak{e} &= [1, 1, \ldots, 1]^T, \\ \mathfrak{p} &= \left[-\frac{1}{2}\sqrt{\frac{(2i-3)(2i+5)}{(2i-1)(2i+3)}}\right]_{i=2}^{p-2}, \\ \mathfrak{d} &= \left[\frac{(2i-3)(2i+1)}{2}\right]_{i=2}^{p}, \end{aligned}$$

cf. [49]. A reordering $\tilde{P}$ of the rows and columns of the matrices F and D gives

$$\tilde{P}F\tilde{P}^T = \begin{bmatrix} F_1 & \mathbf{0} \\ \mathbf{0} & F_2 \end{bmatrix}, \tag{3.3.6}$$

where $F_1 = \text{tridiag}[\mathfrak{e}, \mathfrak{p}_o]$ and $F_2 = \text{tridiag}[\mathfrak{e}, \mathfrak{p}_e]$. Analogously, with the same permutation $\tilde{P}$, one easily derives

$$\tilde{P}D\tilde{P}^T = \begin{bmatrix} D_1 & \mathbf{0} \\ \mathbf{0} & D_2 \end{bmatrix}, \tag{3.3.7}$$

where $D_1 = \text{diag}[\mathfrak{d}_o]$ and $D_2 = \text{diag}[\mathfrak{d}_e]$. The indices o and e denote the odd and even components of the vectors $\mathfrak{p}$ and $\mathfrak{d}$. The matrix $A_{\mathcal{R}_2}$ has some important properties which we summarize in a proposition.

PROPOSITION 3.3. *The following assertions are valid.*

1. *There exists a permutation P of rows and columns such that*

$$PA_{\mathcal{R}_2}P^T = \text{blockdiag}\,[\mathfrak{A}_i]_{i=1}^4$$

   holds.

2. *The matrices $\mathfrak{A}_i$, $i = 1, 2, 3, 4$, are sparse.*

3. *Moreover, each block $\mathfrak{A}_i$ has a 5-point stencil structure.*

4. *The condition number of $\mathfrak{A}_i$ is of order p^2.*

5. *The blocks $\mathfrak{A}_i$ are spectrally equivalent to each other, i.e. $\kappa\left({\mathfrak{A}_i}^{-1}\mathfrak{A}_j\right) = \mathcal{O}(1)$ for $i, j = 1, \ldots, 4$.*

Proof: We note that the four blocks $\mathfrak{A}_i$ correspond to the coefficients of the polynomials $\hat{L}_{2m,2n}$, $\hat{L}_{2m,2n+1}$, $\hat{L}_{2m+1,2n}$, and $\hat{L}_{2m+1,2n+1}$. From (3.3.6) and (3.3.7), we deduce

$$\mathfrak{A}_{2i+j-2} = F_j \otimes D_i + D_j \otimes F_i \quad i, j = 1, 2.$$

Thus, the first three assertions follow immediately from (3.3.4) and (3.3.5). By $\kappa\left({D_1}^{-1}D_2\right) = \mathcal{O}(1)$ which is trivial and $\kappa\left({F_1}^{-1}F_2\right) = \mathcal{O}(1)$, cf. [49], the last assertion follows. The fourth assertion is proved in [49]. □

Similar results are valid for $A_{\mathcal{R}_3}$. We introduce the matrices

$$\mathfrak{B}_{4i+2j+k-6} = F_i \otimes F_j \otimes D_k + F_i \otimes D_j \otimes F_k + D_i \otimes F_j \otimes F_k$$

for $i, j, k = 1, 2$. Using similar arguments as in Proposition 3.3, the next proposition follows.

PROPOSITION 3.4. *There exists a permutation $\hat{P}$ of rows and columns such that*

$$\hat{P}A_{\mathcal{R}_3}\hat{P}^T = \text{blockdiag}\,[\mathfrak{B}_i]_{i=1}^8$$

holds. The blocks $\mathfrak{B}_i$ are spectrally equivalent to each other, i.e. $\kappa\left({\mathfrak{B}_i}^{-1}\mathfrak{B}_j\right) = \mathcal{O}(1)$ for all $i, j = 1, \ldots, 8$.

In the following, we will focus on finding an efficient preconditioner for $\mathfrak{A}_1$, and $\mathfrak{B}_1$. Via Propositions 3.3 and 3.4, the preconditioner for $A_{\mathcal{R}_d}$, $d = 2, 3$ can be constructed. For reasons of simplicity, we assume that p is odd. Furthermore, let $n - 1 = \frac{p-1}{2}$ be the dimension of F_1, and D_1.

3.4 Preconditioner for the element stiffness matrix

3.4.1 Preconditioner of Jensen and Korneev

In [49], Jensen and Korneev have derived a preconditioner for the matrix $A_{\mathcal{R}_2}$, or equivalently, for $\mathfrak{A}_1$. Using $\mathfrak{e} = [1, 1, \ldots, 1]^T$, the matrices

$$D_3 = 4 \operatorname{diag}\left[i^2\right]_{i=1}^{n-1}, \tag{3.4.1}$$

$$T_1 = D_3^{-1} + \frac{1}{2}\operatorname{tridiag}\left[2\mathfrak{e}, -\mathfrak{e}\right], \tag{3.4.2}$$

$$C_1 = D_3 \otimes T_1 + T_1 \otimes D_3 \tag{3.4.3}$$

are introduced. Then, the following lemma holds.

LEMMA 3.5. *The following eigenvalue estimates are valid:*

$$\lambda_{min}\left({D_3}^{-1}D_1\right) \asymp 1, \qquad \lambda_{max}\left({D_3}^{-1}D_1\right) \asymp 1, \tag{3.4.4}$$

$$\lambda_{min}\left({T_1}^{-1}F_1\right) \asymp 1, \qquad \lambda_{max}\left({T_1}^{-1}F_1\right) \asymp 1, \tag{3.4.5}$$

$$\lambda_{min}\left({C_1}^{-1}\mathfrak{A}_1\right) \asymp 1, \qquad \lambda_{max}\left({C_1}^{-1}\mathfrak{A}_1\right) \asymp 1. \tag{3.4.6}$$

Proof: The estimates (3.4.4) are trivial, (3.4.5) are proved in [49], and the assertions (3.4.6) follow by Lemma 2.5 from (3.4.4) and (3.4.5).□
In the matrix C_1, the same matrix entries are nonzero as in $\mathfrak{A}_1$, but the structure of the nonzero elements is simpler. However, a fast solver for C_1 is needed as well as for $\mathfrak{A}_1$.

3.4.2 Modification of the preconditioner in 1D

Now, the preconditioners (3.4.1) and (3.4.2) are modified in several steps. The resulting matrices can be interpreted as stiffness matrices of discretizations of degenerated elliptic problems which will be shown in chapter 4. In a first step, the matrix T_1 is simplified. Let

$$T_2 = \frac{1}{2}\operatorname{tridiag}\left[2\mathfrak{e}, -\mathfrak{e}\right]. \tag{3.4.7}$$

We prove now the following lemma, cf. [10].

LEMMA 3.6. *The eigenvalues of the matrix* ${T_2}^{-\frac{1}{2}}T_1{T_2}^{-\frac{1}{2}}$ *can be estimated by* $\lambda_{min}\left({T_2}^{-1}T_1\right) \geq 1$ *and* $\lambda_{max}\left({T_2}^{-1}T_1\right) \preceq (1 + \log n)$*, where the parameter* $n - 1$ *denotes the dimension of the matrices* T_1 *and* T_2*.*

Proof: The lower eigenvalue estimate is trivial. In order to prove the upper eigenvalue estimate, we use (3.4.2) and (3.4.7). Then,

$$\begin{aligned}\lambda_{max}\left({T_2}^{-1}T_1\right) &= \lambda_{max}\left({T_2}^{-1}(T_2+D_3^{-1})\right)\\ &= 1+\lambda_{max}\left({T_2}^{-1}D_3^{-1}\right) = 1+\lambda_{max}(D_3^{-\frac{1}{2}}T_2^{-1}D_3^{-\frac{1}{2}}).\end{aligned}$$

The matrix

$$H=[h_{ij}]_{i,j=1}^{n-1}=D_3^{-\frac{1}{2}}T_2^{-1}D_3^{-\frac{1}{2}}$$

can be written explicitely, cf. [27]:

$$H=\frac{1}{2n}\begin{bmatrix} n-1 & \frac{n-2}{2} & \frac{n-3}{3} & \frac{n-4}{4} & \cdots & \frac{2}{n-2} & \frac{1}{n-1}\\ \frac{n-2}{2} & \frac{n-2}{2} & \frac{n-3}{3} & \frac{n-4}{4} & \cdots & \frac{2}{n-2} & \frac{1}{n-1}\\ \frac{n-3}{3} & \frac{n-3}{3} & \frac{n-3}{3} & \frac{n-4}{4} & \cdots & \frac{2}{n-2} & \frac{1}{n-1}\\ \vdots & & & & \ddots & & \vdots\\ \frac{2}{n-2} & \frac{2}{n-2} & & \cdots & & \frac{2}{n-2} & \frac{1}{n-1}\\ \frac{1}{n-1} & \frac{1}{n-1} & \frac{1}{n-1} & \cdots & & \frac{1}{n-1} & \frac{1}{n-1}\end{bmatrix}.$$

Therefore, one easily checks $h_{ij}\geq h_{kj}>0$ for $i>k$ and $j=1,\ldots,n-1$. Thus, by the harmonic series, the estimate

$$\begin{aligned}\sup_i\left(\sum_{j=1}^{n-1}h_{ij}\right)=\sum_{j=1}^{n-1}h_{1j} &= \frac{1}{2n}\sum_{j=1}^{n-1}\frac{n-j}{j}\\ &= \sum_{j=1}^{n-1}\frac{1}{2j}-\frac{1}{2n}\sum_{j=1}^{n-1}1\\ &\leq c(1+\log n)\end{aligned}$$

can be concluded. Using the Perron–Frobenius theorem for nonnegative matrices, [31], we obtain

$$\lambda_{max}(H)\leq c(1+\log n)$$

which proves the lemma. □
In a second step, the diagonal matrix D_3 is modified. We define the matrix D_4 by

$$D_4=4\ \mathrm{diag}\left[i^2+\frac{1}{6}\right]_{i=1}^{n-1}. \tag{3.4.8}$$

The next proposition is trivial.

PROPOSITION 3.7. *The eigenvalue estimates* $\lambda_{min}\left({D_4}^{-1}D_3\right)=\frac{6}{7}$ *and* $\lambda_{max}\left({D_4}^{-1}D_3\right)<1$ *are valid.*

Now, the matrix D_3 is changed in another way. Let

$$D_5 = \text{tridiag}[\mathfrak{b}, \mathfrak{a}], \tag{3.4.9}$$

where

$$\begin{aligned} \mathfrak{a} &= \left[i^2 + i + \frac{3}{10} \right]_{i=1}^{n-2}, \\ \mathfrak{b} &= \left[4i^2 + \frac{2}{5} \right]_{i=1}^{n-1}. \end{aligned}$$

By the following lemma, the condition number of the matrix $D_5^{-1} D_3$ is bounded by a constant independent of n.

LEMMA 3.8. *The eigenvalue estimates* $\lambda_{min}\left({D_5}^{-1} D_3\right) \asymp 1$ *and* $\lambda_{max}\left({D_5}^{-1} D_3\right) \asymp 1$ *hold.*

Proof: An easy calculation shows

$$H_1 = [h_{ij}^{(1)}]_{i,j=1}^{n-1} = D_3^{-\frac{1}{2}} D_5 D_3^{-\frac{1}{2}} = \text{tridiag}[\mathfrak{g}, \mathfrak{f}],$$

where

$$\begin{aligned} \mathfrak{f} &= \left[\frac{1}{4} + \frac{3}{40(i^2+i)} \right]_{i=1}^{n-2}, \\ \mathfrak{g} &= \left[1 + \frac{1}{10i^2} \right]_{i=1}^{n-1}. \end{aligned}$$

Taking Gerschgorins-disks, [34], we obtain the estimate

$$\min_i \left(h_{ii}^{(1)} - \sum_{j \neq i} \mid h_{ij}^{(1)} \mid \right) \leq \lambda_{min}(H_1) \leq \lambda_{max}(H_1) \leq \max_i \left(h_{ii}^{(1)} + \sum_{j \neq i} \mid h_{ij}^{(1)} \mid \right).$$

Using the structure of $\mathfrak{f}$ and $\mathfrak{g}$, we can conclude

$$\begin{aligned} \min_i \left(h_{ii}^{(1)} - \sum_{j \neq i} \mid h_{ij}^{(1)} \mid \right) &\geq \frac{1}{2}, \\ \max_i \left(h_{ii}^{(1)} + \sum_{j \neq i} \mid h_{ij}^{(1)} \mid \right) &\leq \frac{63}{40}. \end{aligned}$$

Hence, the assertions follow.□
Recall that the inverse of the matrix D_3 is required for the definition of the matrix T_1 (3.4.3). Now, we introduce a tridiagonal matrix D_6 from which we will show that $\kappa\left(D_3 D_6\right) \leq c$. Let

$$D_6 = \text{tridiag}[\mathfrak{h}, \mathfrak{r}], \tag{3.4.10}$$

where

$$\begin{aligned}\mathfrak{h} &= \frac{1}{2}\left[(j-1)\ln(j-1) - (j+1)\ln(j+1) + 2\ln(j) + 2\right]_{j=1}^{n-1},\\ \mathfrak{k} &= \frac{1}{4}\left[-2 + (2j+1)\ln(j+1) - (2j+1)\ln(j)\right]_{j=1}^{n-2}.\end{aligned}$$

For reasons of simplicity, the undefined value "$0\ln 0$" is 0 by definition. It will be shown in the next chapter that the matrix D_6 can be interpreted as a weighted mass-matrix.
The following result is valid.

LEMMA 3.9. *The condition number of D_3D_6 is bounded by a constant independent of n, i.e. $\kappa(D_3D_6) \le c$.*

Proof: The proof is similar to the proof of Lemma 3.8. More precisely, we determine the entries of the symmetric tridiagonal matrix

$$H_2 = [h_{ij}^{(2)}]_{i,j=1}^{n-1} = D_3^{\frac{1}{2}} D_6 D_3^{\frac{1}{2}}$$

and take Gerschgorin disks. Then, one easily checks

$$\begin{aligned}h_{jj}^{(2)} &= 4j^2 + 4j^2\ln j + 2j^2(j-1)\ln(j-1) - 2j^2(j+1)\ln(j+1),\\ h_{j+1,j}^{(2)} = h_{j,j+1}^{(2)} &= (2j+1)j(j+1)\ln(j+1) - (2j+1)j(j+1)\ln j - 2j(j+1).\end{aligned}$$

One easily verifies that $h_{ij} \ge 0$ for all $i, j \in \mathbb{N}$. Moreover, we obtain

$$\begin{aligned}h_{j,j-1}^{(2)} + h_{jj}^{(2)} + h_{j,j+1}^{(2)} &= 2\left(j^2\ln\frac{j^2-1}{j^2} + j\ln\frac{j+1}{j-1}\right),\\ -h_{j,j-1}^{(2)} + h_{jj}^{(2)} - h_{j,j+1}^{(2)} &= 2\left(5j^2\ln\frac{j^2}{j^2-1} + 8j^2 + j\ln\frac{j-1}{j+1} + 4j^2\ln\frac{j-1}{j+1}\right)\end{aligned}$$

for $j \ge 2$. The function $f : (1,\infty) \mapsto \mathbb{R}$,

$$f(x) = 2\left(x^2\ln\frac{x^2-1}{x^2} + x\ln\frac{x+1}{x-1}\right)$$

is monotonic decreasing for $x \ge 2$. It attains its maximum on $[2,\infty)$ at $x = 2$, where

$$\max_{x\in[2,\infty)} f(x) = f(2) = 12\ln 3 - 16\ln 2. \tag{3.4.11}$$

The function $g : (1,\infty) \mapsto \mathbb{R}$,

$$g(x) = 2\left(5x^2\ln\frac{x^2}{x^2-1} + 8x^2 + x\ln\frac{x-1}{x+1} + 4x^2\ln\frac{x-1}{x+1}\right)$$

is monotonic decreasing for $x \geq 2$ and satisfies

$$\inf_{x\in[2,\infty)} g(x) = \lim_{x\to\infty} g(x) = \frac{2}{3}, \tag{3.4.12}$$

which is its infimum on the interval $[2, \infty)$. Moreover, by a direct calculation, the relations $h_{11}^{(2)} = 4 - 4\ln 2$ and $h_{12}^{(2)} = 6\ln 2 - 4$ are valid. Thus,

$$h_{11}^{(2)} + h_{12}^{(2)} = 2\ln 2, \tag{3.4.13}$$
$$h_{11}^{(2)} - h_{12}^{(2)} = 8 - 10\ln 2. \tag{3.4.14}$$

By (3.4.11) and (3.4.13), the lower eigenvalue estimate

$$\lambda_{min}\left(D_3 D_6\right) \leq 12\ln 3 - 16\ln 2$$

follows. By (3.4.12) and (3.4.14), one obtains the upper eigenvalue estimate

$$\lambda_{max}\left(D_3 D_6\right) \leq \frac{2}{3}$$

which proves the lemma. □
Now, we introduce the matrix

$$T_3 = D_6 + T_2. \tag{3.4.15}$$

Then by Lemma 3.9, the following conclusion can be drawn.

COROLLARY 3.10. *The matrix $T_1 = D_3^{-1} + T_2$ is spectrally equivalent to the matrix T_3, i.e. $\kappa\left({T_1}^{-1} T_3\right) \leq c$.*

Proof: Use Lemma 3.9 and the fact that D_3, D_6 and T_2 are symmetric and positive definite matrices. □

3.4.3 Modification of the preconditioner in 2D and 3D

Via tensor product and by the relations (3.4.1) for D_3, (3.4.8) for D_4, (3.4.9) for D_5, (3.4.7) for T_2, and (3.4.15) for T_3, the matrices

$$C_2 = D_5 \otimes T_3 + T_3 \otimes D_5, \tag{3.4.16}$$
$$C_3 = D_3 \otimes T_2 + T_2 \otimes D_3, \tag{3.4.17}$$
$$C_4 = D_4 \otimes T_2 + T_2 \otimes D_4, \tag{3.4.18}$$
$$C_5 = D_5 \otimes T_2 + T_2 \otimes D_5 \tag{3.4.19}$$

are introduced. Then, the following theorem holds.

THEOREM 3.11. *For $i = 3, 4, 5$, the eigenvalue estimates*

$$\lambda_{min}\left(C_i{}^{-1}\mathfrak{A}_1\right) \asymp 1 \quad \textit{and} \quad \lambda_{max}\left(C_i{}^{-1}\mathfrak{A}_1\right) \preceq (1 + \log n)$$

are valid. Moreover, the condition number of the matrix $C_2^{-1}\mathfrak{A}_1$ is bounded by a constant independent of n, i.e.

$$\kappa\left(C_2{}^{-1}\mathfrak{A}_1\right) \asymp 1.$$

Proof: Note that D_3, D_4, D_5, T_3, and T_2 are symmetric and positive definite and apply Lemma 2.5 to the matrices C_3, C_4, C_5, and C_2. By Lemma 3.5, Lemma 3.6, Lemma 3.8, Corollary 3.10, and Proposition 3.7, the assertions follow. □
In the same way, with D_3 (3.4.1), D_5 (3.4.9), T_1 (3.4.2), T_2 (3.4.7), and T_3 (3.4.15), the matrices

$$C_6 = D_3 \otimes T_1 \otimes T_1 + T_1 \otimes D_3 \otimes T_1 + D_3 \otimes T_1 \otimes T_1, \tag{3.4.20}$$
$$C_7 = D_3 \otimes T_2 \otimes T_2 + T_2 \otimes D_3 \otimes T_2 + D_3 \otimes T_2 \otimes T_2, \tag{3.4.21}$$
$$C_8 = D_5 \otimes T_2 \otimes T_2 + T_2 \otimes D_5 \otimes T_2 + D_5 \otimes T_2 \otimes T_2, \tag{3.4.22}$$
$$C_9 = D_5 \otimes T_3 \otimes T_3 + T_3 \otimes D_5 \otimes T_3 + D_5 \otimes T_3 \otimes T_3 \tag{3.4.23}$$

are defined. By the same arguments as in Theorem 3.11, the next theorem can be proved.

THEOREM 3.12. *The following eigenvalue estimates are valid:*

- $\lambda_{min}\left(C_i{}^{-1}\mathfrak{B}_1\right) \asymp 1$ *for* $i = 6, 7, 8, 9$,
- $\lambda_{max}\left(C_i{}^{-1}\mathfrak{B}_1\right) \asymp 1$ *for* $i = 6, 9$,
- $\lambda_{max}\left(C_i{}^{-1}\mathfrak{B}_1\right) \preceq (1 + \log n)^2$ *for* $i = 7, 8$.

4 Interpretation of the preconditioners

In the previous chapter, several preconditioners for the matrices F_1 and D_1, cf. (3.3.6) and (3.3.7), are derived. In this chapter, we show that these preconditioners can be interpreted as matrices resulting from the discretization of several auxiliary problems. We distinguish between the three cases 1D, 2D and 3D, and approximations by finite elements or finite differences.

4.1 The one-dimensional case

4.1.1 Finite differences

Consider the following problem: Find u such that

$$\begin{aligned} -\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} + \frac{1}{x^2}u + x^2 u &= g \quad \text{for } x \in (0,1), \\ u(0) = u(1) &= 0 \end{aligned} \tag{4.1.1}$$

holds. Problem (4.1.1) is discretized by finite differences. Let k be the level number, and let $n = 2^k$. Moreover, let

$$x_j^k = \frac{j}{n}, \quad j = 0, \ldots, n,$$

be a set of grid points in the interval [0,1]. On the grid $\{x_j^k\}_{j=1}^{n-1}$, let u_j^k be the (approximated) value of u in the point x_j^k. The terms of (4.1.1) at x_j^k are approximated by

$$\begin{aligned} -\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} &\approx \frac{-u_{j-1}^k + 2u_j^k - u_{j+1}^k}{h^2}, \\ x^2 u &\approx u_j^k \frac{j^2}{n^2} = h^2 j^2 u_j^k, \\ \frac{1}{x^2}u &\approx u_j^k \frac{n^2}{j^2} = \frac{1}{h^2 j^2} u_j^k, \end{aligned}$$

where $h = \frac{1}{n}$. Then, the finite difference approximation of (4.1.1) can be rewritten as

$$\begin{aligned} \frac{1}{h^2}\left(-u_{j-1}^k + 2u_j^k - u_{j+1}^k\right) + \frac{1}{h^2}\frac{u_j^k}{j^2} + h^2 j^2 u_j^k &= g(u_j^k), \quad j = 1, \ldots, n-1, \\ u_0^k &= 0, \\ u_n^k &= 0. \end{aligned} \tag{4.1.2}$$

This problem is equivalent to solving

$$\frac{2}{h^2}\left(T_2 + 2D_3^{-1}\right) + \frac{h^2}{4}D_3 = \left(\frac{2}{h^2}T_2 + \frac{4}{h^2}D_3^{-1} + \frac{h^2}{4}D_3\right)\underline{u} = \underline{g},$$

where $\underline{u} = [u_j^k]_{j=1}^{n-1}$ and $\underline{g} = [g(u_j^k)]_{j=1}^{n-1}$ with the matrices T_2 (3.4.7) and D_3 (3.4.1).

4.1.2 Finite elements

Consider now problem (4.1.1) in the weak formulation.
Find $u \in H_0^1((0,1)) \cap L^2_{\omega=x}((0,1)) \cap L^2_{\omega=x^{-1}}((0,1))$ such that

$$a_1(u,v) = a_s(u,v) + a_{\overline{m}}(u,v) + a_m(u,v) = \langle g, v\rangle \tag{4.1.3}$$

holds for all $v \in H_0^1((0,1)) \cap L^2_{\omega=x}((0,1)) \cap L^2_{\omega=x^{-1}}((0,1))$. The bilinear forms $a_s(\cdot,\cdot)$, $a_{\overline{m}}(\cdot,\cdot)$ and $a_m(\cdot,\cdot)$ are defined as

$$\begin{aligned} a_s(u,v) &= \int_0^1 u'(x)v'(x)\,\mathrm{d}x, \\ a_{\overline{m}}(u,v) &= \int_0^1 x^{-2}u(x)v(x)\,\mathrm{d}x, \\ a_m(u,v) &= \int_0^1 x^2 u(x)v(x)\,\mathrm{d}x. \end{aligned}$$

This one-dimensional problem (4.1.3) is discretized by linear finite elements on the equidistant mesh

$$T_k = \bigcup_{i=0}^{n-1} \tau_i^k,$$

where

$$\tau_i^k = \left(\frac{i}{n}, \frac{i+1}{n}\right).$$

As in the previous subsection, the parameter k denotes the level number. On this mesh, we introduce the one-dimensional hat-functions

$$\phi_i^{(1,k)}(x) = \begin{cases} nx - (i-1) & \text{on} \quad \tau_{i-1}^k \\ (i+1) - nx & \text{on} \quad \tau_i^k \\ 0 & \text{else} \end{cases}, \quad i = 1,\ldots,n-1, \tag{4.1.4}$$

where $n = 2^k$. Let $\mathbb{V}_k^{(1)} = \operatorname{span}\{\phi_i^{(1,k)}\}_{i=1}^{n-1}$ be the corresponding finite element space. Then, the Galerkin projection of (4.1.3) onto $\mathbb{V}_k^{(1)}$ is:
Find $u^k \in \mathbb{V}_k^{(1)}$ such that

$$a_1(u^k, v^k) = \langle g, v^k\rangle \quad \forall v^k \in \mathbb{V}_k^{(1)}. \tag{4.1.5}$$

Then using (3.4.7), we obtain

$$\left[a_s(\phi_j^{(1,k)}, \phi_i^{(1,k)})\right]_{i,j=1}^{n-1} = 2n\, T_2 = n\ \mathrm{tridiag}[2\mathfrak{e}, -\mathfrak{e}]. \tag{4.1.6}$$

Moreover, an easy calculation shows

$$\left[a_{\overline{m}}(\phi_j^{(1,k)}, \phi_i^{(1,k)})\right]_{i,j=1}^{n-1} = 4nD_6 \tag{4.1.7}$$

and

$$\left[a_m(\phi_j^{(1,k)}, \phi_i^{(1,k)})\right]_{i,j=1}^{n-1} = \frac{1}{6n^3} D_5. \tag{4.1.8}$$

By (4.1.6), (4.1.7), and (3.4.15), one checks

$$\left[2a_s(\phi_j^{(1,k)}, \phi_i^{(1,k)}) + a_{\overline{m}}(\phi_j^{(1,k)}, \phi_i^{(1,k)})\right]_{i,j=1}^{n-1} = 4nT_3. \tag{4.1.9}$$

Hence, interpretations of the matrices $T_2 \in \mathbb{R}^{n-1\times n-1}$ (3.4.7), $T_3 \in \mathbb{R}^{n-1\times n-1}$ (3.4.15), $D_5 \in \mathbb{R}^{n-1\times n-1}$ (3.4.9) and $D_6 \in \mathbb{R}^{n-1\times n-1}$ (3.4.10) have been given.

4.2 The two-dimensional case

4.2.1 Finite differences

We consider the following second order problem: Find u such that

$$\begin{aligned} -2\,(y^2 u_{xx} + x^2 u_{yy}) &= g \quad \text{in} \quad \Omega = (0,1)^2, \\ u &= 0 \quad \text{on} \quad \partial\Omega. \end{aligned} \tag{4.2.1}$$

Problem (4.2.1) is solved approximately by finite differences on the grid shown in Figure 4.1. The approximation in $(\frac{i}{n}, \frac{j}{n})$ of u is denoted by $u_{i,j}$. The second order derivatives are approximated by the usual central difference quotient, i.e.

$$\begin{aligned} y^2 u_{xx}\left(\frac{i}{n}, \frac{j}{n}\right) &\approx j^2(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}), \\ x^2 u_{yy}\left(\frac{i}{n}, \frac{j}{n}\right) &\approx i^2(u_{i,j-1} - 2u_{i,j} + u_{i,j+1}). \end{aligned}$$

We insert the boundary condition and sort the unknowns in the order $u_{1,1}$, $u_{1,2}$, ..., $u_{1,n-1}$, $u_{2,1}$, ..., $u_{n-1,n-1}$. Then, one obtains by tensor product arguments and the results of subsection 4.1.1 that C_3, which is defined in (3.4.17), is the system matrix for the resulting system of linear algebraic equations. Therefore, the following lemma has been proved.

LEMMA 4.1. *The discretization of (4.2.1) on a uniform grid by finite differences yields to a system of linear algebraic equations of the type* $C_3\underline{u} = \underline{g}$.

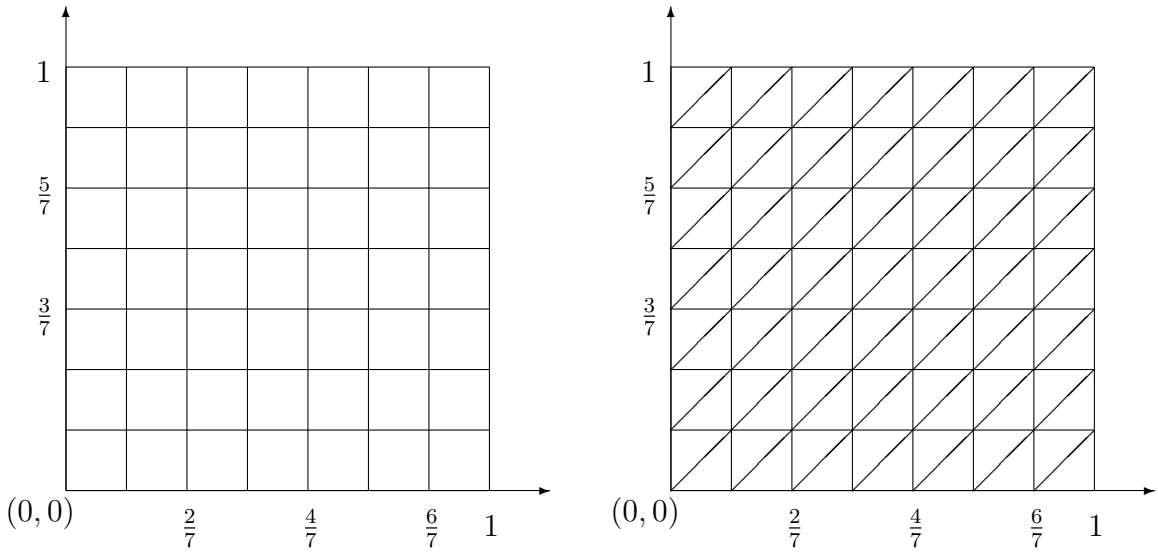

Figure 4.1: Mesh for h-version (left), grid (right).

Considering the matrix C_1 (3.4.3), a similar result as in Lemma 4.1 can be shown, cf. [53]. We state this result as a remark.

REMARK 4.2. *The discretization of the problem*

$$\begin{aligned} -2\left(y^2 u_{xx} + x^2 u_{yy}\right) + \left(\frac{x^2}{y^2} + \frac{y^2}{x^2}\right) u &= g \quad \text{in} \quad \Omega = (0,1)^2, \\ u &= 0 \quad \text{on} \quad \partial\Omega \end{aligned} \tag{4.2.2}$$

as above leads to the linear system $C_1 \underline{u} = \underline{g}$ *with* C_1 *defined in (3.4.3).*

Hence, interpretations of the system matrices C_1 (3.4.3) and C_3 (3.4.17) have been found.

4.2.2 Linear elements on triangles

Consider the following Dirichlet problem: Find $u \in H^1_{0,\omega}(\Omega)$ such that

$$a(u,v) := \int_\Omega \left((\omega(y))^2 u_x v_x + (\omega(x))^2 u_y v_y\right) \, \mathrm{d}x\mathrm{d}y = \int_\Omega gv \, \mathrm{d}x\mathrm{d}y =: \langle g, v\rangle \tag{4.2.3}$$

for all $v \in H^1_{0,\omega}(\Omega)$ holds. The domain Ω is the unit square $(0,1)^2$ and

$$H^1_{0,\omega}(\Omega) = \{u \in L^2(\Omega), \omega(x)u_y, \omega(y)u_x \in L^2(\Omega), u\,|_{\partial\Omega} = 0\}$$

with $\omega(\xi) = \xi$. We discretize problem (4.2.3) by finite elements. For this purpose, some notation is introduced. Let k be the level of approximation and $n = 2^k$. Let $x^k_{ij} = (\frac{i}{n}, \frac{j}{n})$, where $i, j =$

$0, \ldots, n$. The domain Ω is divided into congruent, isosceles, right-angled triangles $\tau_{ij}^{s,k}$, where $0 \leq i, j < n$ and $s = 1, 2$, see Figure 4.1. The triangle $\tau_{ij}^{1,k}$ has the three vertices $x_{ij}^k, x_{i+1,j+1}^k$ and $x_{i,j+1}^k$, $\tau_{ij}^{2,k}$ has the three vertices $x_{ij}^k, x_{i+1,j+1}^k$ and $x_{i+1,j}^k$, see Figure 4.2.

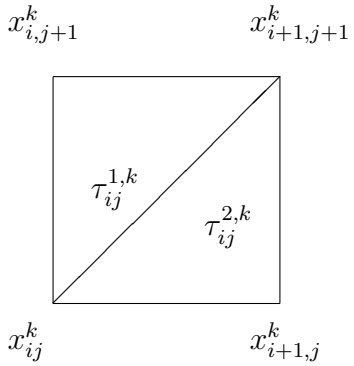

Figure 4.2: Introduction of the geometrical notation of a macro-element $\mathcal{E}_{ij}^k$.

Furthermore, let $\mathcal{E}_{ij}^k = \overline{\tau_{ij}^{1,k} \cup \tau_{ij}^{2,k}}$ be the macro-element

$$\left[\frac{i}{n}, \frac{i+1}{n}\right] \times \left[\frac{j}{n}, \frac{j+1}{n}\right].$$

Piecewise linear finite elements are used on the mesh

$$T_k = \{\tau_{ij}^{s,k}\}_{i=0,j=0,s=1}^{n-1,n-1,2}.$$

The subspace of piecewise linear functions ϕ_{ij}^k with

$$\phi_{ij}^k \in H_0^1(\Omega),\ \phi_{ij}^k\,|_{\tau_{lm}^{s,k}} \in \mathbb{P}^1(\tau_{lm}^{s,k})$$

is denoted by $\mathbb{V}_k$, where $\mathbb{P}^1$ is the space of polynomials of degree ≤ 1. A basis of $\mathbb{V}_k$ is the system of the usual hat-functions $\{\phi_{ij}^k\}_{i,j=1}^{n-1}$ uniquely defined by

$$\phi_{ij}^k(x_{lm}^k) = \delta_{il}\delta_{jm} \tag{4.2.4}$$

and $\phi_{ij}^k \in \mathbb{V}_k$, where δ_{il} is the Kronecker delta. Now, we can formulate the discretized problem: Find $u^k \in \mathbb{V}_k$ such that

$$a(u^k, v^k) = \langle g, v^k \rangle \quad \forall v^k \in \mathbb{V}_k \tag{4.2.5}$$

holds. Problem (4.2.5) is equivalent to solving the system of linear algebraic equations

$$K_k \underline{u}_k = \underline{g}_k, \tag{4.2.6}$$

where

$$\begin{aligned} K_k &= \left[a(\phi_{lm}^k, \phi_{ij}^k)\right]_{i,j,l,m=1}^{n-1}, \\ \underline{u}_k &= \left[u_{ij}\right]_{i,j=1}^{n-1}, \\ \underline{g}_k &= \left[\langle g, \phi_{lm}^k \rangle\right]_{l,m=1}^{n-1}. \end{aligned}$$

Then, $u^k = \sum_{i,j=1}^{n-1} u_{ij}\phi_{ij}^k$ is the solution of (4.2.5).

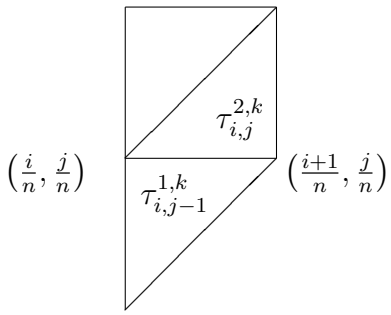

Figure 4.3: Sketch for the computation of the matrix entry between two adjacent nodes.

We determine now $a(\phi_{ij}^k, \phi_{i+1,j}^k)$. One obtains by a simple integration, cf. Figure 4.3.

$$\begin{aligned} a(\phi_{ij}^k, \phi_{i+1,j}^k) &= \int_{\tau_{i,j-1}^{1,k}} \begin{bmatrix} -n \\ n \end{bmatrix}^T \begin{bmatrix} y^2 & 0 \\ 0 & x^2 \end{bmatrix} \begin{bmatrix} n \\ 0 \end{bmatrix} \mathrm{d}(x,y) \\ &\quad + \int_{\tau_{ij}^{2,k}} \begin{bmatrix} -n \\ 0 \end{bmatrix}^T \begin{bmatrix} y^2 & 0 \\ 0 & x^2 \end{bmatrix} \begin{bmatrix} n \\ -n \end{bmatrix} \mathrm{d}(x,y) \\ &= -n^2 \int_{\tau_{i,j-1}^{1,k} \cup \tau_{ij}^{2,k}} y^2 \, \mathrm{d}(x,y) \\ &= -n^2 \int_{\frac{j-1}{n}}^{\frac{j}{n}} \int_{\frac{i}{n}}^{y+\frac{i-j+1}{n}} y^2 \, \mathrm{d}x\mathrm{d}y - n^2 \int_{\frac{j}{n}}^{\frac{j+1}{n}} \int_{y+\frac{i-j}{n}}^{\frac{i+1}{n}} y^2 \, \mathrm{d}x\mathrm{d}y \\ &= -\frac{1}{n^2}\left(\frac{j^2}{2} - \frac{j}{3} + \frac{1}{12}\right) - \frac{1}{n^2}\left(\frac{j^2}{2} + \frac{j}{3} + \frac{1}{12}\right) \\ &= -\frac{1}{n^2}\left(\frac{1}{6} + j^2\right), \end{aligned} \tag{4.2.7}$$

where $n > i, j$ and $j > 0$, but $i \geq 0$. By symmetry of the differential operator in (4.2.3) with respect to the variables x and y, it follows

$$a(\phi_{ij}^k, \phi_{i,j+1}^k) = -\frac{1}{n^2}\left(\frac{1}{6} + i^2\right), \tag{4.2.8}$$

where $i > 0$ and $j \geq 0$ and

$$\begin{aligned} a(\phi_{ij}^k, \phi_{ij}^k) &= -(a(\phi_{ij}^k, \phi_{i+1,j}^k) + a(\phi_{ij}^k, \phi_{i,j+1}^k) + a(\phi_{ij}^k, \phi_{i,j-1}^k) + a(\phi_{ij}^k, \phi_{i-1,j}^k)) \\ &= \frac{1}{n^2}\left(2i^2 + 2j^2 + \frac{2}{3}\right). \end{aligned}$$

Inserting the boundary condition and using (3.4.18), we arrive at

$$K_k = \frac{1}{2n^2} C_4 \tag{4.2.9}$$

after a proper permutation of the unknowns. Thus, an interpretation for the matrix C_4 (3.4.18) has been found. Thus, the following lemma has been proved.

LEMMA 4.3. *The discretization of (4.2.3) by piecewise linear finite elements on the mesh of Figure 4.1 is equivalent to the system solve of linear algebraic equations (4.2.6), where* $K_k = \frac{1}{2n^2} C_4$.

4.2.3 Bilinear elements on quadrilaterals

As in the previous subsection, consider problem (4.2.3): Find $u \in H^1_{0,\omega}(\Omega)$ such that

$$a(u, v) = \int_\Omega \left((\omega(y))^2 u_x v_x + (\omega(x))^2 u_y v_y\right) \, \mathrm{d}x\mathrm{d}y = \int_\Omega gv \, \mathrm{d}x\mathrm{d}y = \langle g, v\rangle \tag{4.2.10}$$

for all $v \in H^1_{0,\omega}(\Omega)$, where the weight function ω satisfies $\omega(\xi) = \xi$. The domain Ω is the unit square $(0,1)^2$. We want to find an approximate solution of (4.2.10) using bilinear finite elements on quadrilaterals. The following notations are needed. As in subsection 4.2.2, let k be the level of approximation and $n = 2^k$. Let $x_{ij}^k = (\frac{i}{n}, \frac{j}{n})$, where $i, j = 0, \ldots, n$. The domain Ω is divided into congruent squares $\mathcal{E}_{ij}^k = \overline{\tau_{ij}^{1,k} \cup \tau_{ij}^{2,k}}$, i.e.

$$\mathcal{E}_{ij}^k = \left[\frac{i}{n}, \frac{i+1}{n}\right] \times \left[\frac{j}{n}, \frac{j+1}{n}\right].$$

On the mesh of squares

$$\mathcal{E}^k = \{\mathcal{E}_{ij}^k\}_{i,j=0}^{n-1},$$

the piecewise bilinear shape functions $\phi_{ij}^{b,k}$ are introduced as tensor products of the one-dimensional functions $\phi_i^{(1,k)}$, cf. (4.1.4),

$$\phi_{ij}^{b,k}(x, y) = \phi_i^{(1,k)}(x)\phi_j^{(1,k)}(y) \quad \text{for} \quad i, j = 1, \ldots, n-1.$$

Set $\mathbb{V}_k^{(b)} = \mathrm{span}\{\phi_{ij}^{b,k}\}_{i,j=1}^{n-1}$. Now, the discrete problem can be formulated. Find $u^k \in \mathbb{V}_k^{(b)}$ such that

$$a(u^k, v^k) = \langle g, v^k\rangle \quad \forall v^k \in \mathbb{V}_k^{(b)} \tag{4.2.11}$$

holds. Problem (4.2.11) is equivalent to solving

$$K_{b,k}\underline{u}_b = \underline{g}_b, \tag{4.2.12}$$

where

$$\begin{aligned} K_{b,k} &= \left[a(\phi_{lm}^{b,k}, \phi_{ij}^{b,k})\right]_{i,j,l,m=1}^{n-1}, \\ \underline{u}_b &= \left[u_{ij}^b\right]_{i,j=1}^{n-1}, \\ \underline{g}_b &= \left[\langle g, \phi_{lm}^{b,k}\rangle\right]_{l,m=1}^{n-1}. \end{aligned}$$

Then, $u^k = \sum_{i,j=1}^{n-1} u_{ij}^b \phi_{ij}^{b,k}$ is the solution of (4.2.11). From (4.1.6), (4.1.8) and (3.4.19), one can conclude

$$\begin{aligned} K_{b,k} &= \frac{2n}{6n^3}\left(T_2 \otimes D_5 + D_5 \otimes T_2\right), \\ &= \frac{1}{3n^2} C_5. \end{aligned} \tag{4.2.13}$$

Thus, the following lemma is valid.

LEMMA 4.4. *The discretization of problem (4.2.10) by bilinear elements on the mesh* $\left\{\mathcal{E}_{ij}^k\right\}_{i,j=0}^{n-1}$ *is equivalent to solving the system of linear algebraic equations* $C_5\underline{u}_b = \underline{g}_b$.

Moreover, we consider the following discrete problem. Find $u^k \in \mathbb{V}_k$ such that

$$a_2(u^k, v^k) := 2a(u^k, v^k) + \int_\Omega \left(\frac{\omega^2(x)}{\omega^2(y)} + \frac{\omega^2(y)}{\omega^2(x)}\right) u^k v^k = \langle g, v^k\rangle \tag{4.2.14}$$

holds for all $v^k \in \mathbb{V}_k$, where $a(\cdot,\cdot)$ is the bilinear form defined in (4.2.10) and $\omega(\xi) = \xi$. With the same arguments as in the proof of Lemma 4.4, the following result can be shown.

LEMMA 4.5. *Let* C_2 *be defined in (3.4.16). Then, the discretization of problem (4.2.14) by bilinear elements on the mesh* $\left\{\mathcal{E}_{ij}^k\right\}_{i,j=0}^{n-1}$ *is equivalent to solving the system of linear algebraic equations* $C_2\underline{u}_b = \underline{g}_b$.

4.2.4 Improvement for rectangular elements

In (3.3.1), let us assume that the domain $\mathcal{R}_2 = (-1,1)^2$ is replaced by a rectangle, i.e. $\mathcal{R}_2^{a,b} = (-a,a)\times(-b,b)$, where $a, b > 0$. The discretization by the p-version of the finite element method using only one element $\mathcal{R}_2^{a,b}$ leads to a system of the type $K_{a,b}\underline{x} = \underline{y}$, where

$$K_{a,b} = \frac{a}{b}(F \otimes D) + \frac{b}{a}(D \otimes F)$$

with the matrices F and D defined via relation (3.3.5). The matrices C_3 (3.4.17), C_4 (3.4.18), and C_5 (3.4.19) can be used as preconditioner for $K_{a,b}$. However, all eigenvalue estimates will depend on the geometric parameters a and b. Thus, by a simple scaling, new matrices are developed such that the estimates for the eigenvalues do not depend on the parameters a and b. Similar as in Proposition 3.3, the relation

$$K_{a,b} = P \text{ blockdiag } [\mathfrak{A}_{i,a,b}]_{i=1}^{4} P^T$$

holds with the same permutation matrix P, and

$$\mathfrak{A}_{2i+j-2,a,b} = \frac{a}{b}(F_i \otimes D_j) + \frac{b}{a}(D_i \otimes F_j) \quad i,j = 1,2.$$

Instead of (4.2.1), we consider the boundary value problem

$$\begin{aligned} -\frac{a}{b}y^2 u_{xx} - \frac{b}{a}x^2 u_{yy} &= g \quad \text{in} \quad \Omega = (0,1)^2, \\ u &= 0 \quad \text{on} \quad \partial\Omega. \end{aligned} \tag{4.2.15}$$

The discretization of (4.2.15) by finite differences, linear or bilinear elements as described in the subsections 4.2.1–4.2.3 yields to systems of linear algebraic equations with matrices $C_{3,a,b}$, $C_{4,a,b}$ and $C_{5,a,b}$, where

$$C_{i,a,b} = \frac{a}{b}(T_2 \otimes D_i) + \frac{b}{a}(D_i \otimes T_2) \quad i = 3,4,5.$$

Now, we are able to formulate the next lemma.

LEMMA 4.6. *The condition number* $\kappa\left({C_{j,a,b}}^{-1}\mathfrak{A}_{i,a,b}\right)$ *grows as* $(1+\log n)$ *for* $j = 3,4,5$ *and* $i = 1,2,3,4$*, where the constants do not depend on the parameters* a *and* b*.*

Proof: The assertions follow by Lemma 2.5. □

4.3 The three-dimensional case

Consider the fourth order boundary value problems

$$\begin{aligned} z^2 u_{xxyy} + y^2 u_{xxzz} + x^2 u_{yyzz} &= g \quad \text{in} \quad \Omega_3 = (0,1)^3, \\ u &= 0 \quad \text{on} \quad \partial\Omega_3 \end{aligned} \tag{4.3.1}$$

and

$$\begin{aligned} 4\left(z^2 u_{xxyy} + y^2 u_{xxzz} + x^2 u_{yyzz}\right) \\ -2\left(\frac{y^2}{z^2} + \frac{z^2}{y^2}\right) u_{xx} - 2\left(\frac{x^2}{z^2} + \frac{z^2}{x^2}\right) u_{yy} - 2\left(\frac{x^2}{y^2} + \frac{y^2}{x^2}\right) u_{zz} \\ + \left(\frac{x^2}{y^2 z^2} + \frac{y^2}{x^2 z^2} + \frac{z^2}{x^2 y^2}\right) u &= g \quad \text{in} \quad \Omega_3 = (0,1)^3, \\ u &= 0 \quad \text{on} \quad \partial\Omega_3. \end{aligned} \tag{4.3.2}$$

Note that the differential operators in (4.3.1) and (4.3.2) do not have a term of a pure fourth derivative, there are mixed terms only. We will discretize these problems by finite differences or trilinear finite elements on hexahedrons.

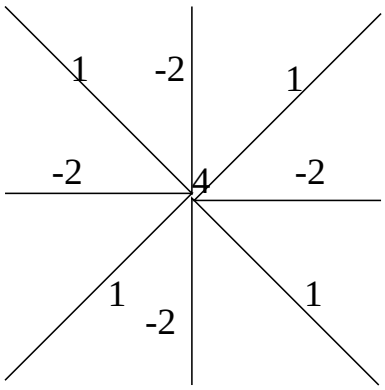

Figure 4.4: Stencil for discretization of u_{xxyy}.

4.3.1 Finite differences

Problems (4.3.1) and (4.3.2) are discretized by the method of finite differences on the equidistant grid $\left\{\frac{1}{n}(i,j,l)\right\}_{i,j,l=1}^{n-1}$, where k denotes the level number and $n = 2^k$. Let $u_{i,j,l}$ be the approximation of u in the point $\frac{1}{n}(i,j,l)$. The mixed fourth order derivatives are discretized by the stencil of Figure 4.4, e.g. for u_{xxyy}

$$\begin{aligned} z^2 u_{xxyy}\left(\frac{i}{n},\frac{j}{n},\frac{l}{n}\right) \quad \approx \quad & n^2 l^2 (4u_{i,j,l} - 2u_{i,j-1,l} - 2u_{i,j+1,l} - 2u_{i-1,j,l} - 2u_{i+1,j,l} \\ & + u_{i-1,j-1,l} + u_{i+1,j-1,l} + u_{i-1,j+1,l} + u_{i+1,j+1,l}), \end{aligned}$$

the second order derivatives of (4.3.2) by the usual central differential quotient.

LEMMA 4.7. *This approximation of (4.3.1) is equivalent to solving the system of linear algebraic equations $C_7\underline{u} = \underline{g}$ with C_7 defined in (3.4.21). Moreover, the finite difference approximation of problem (4.3.2) is equivalent to solving $C_6\underline{u} = \underline{g}$, where C_6 is defined via relation (3.4.20).*

Proof: The left hand side $z^2u_{xxyy} + y^2u_{xxzz} + x^2u_{yyzz}$ of the partial differential equation of problem (4.3.1) can be written into the form

$$\left(\frac{\partial^2}{\partial x^2}\frac{\partial^2}{\partial y^2}z^2 + \frac{\partial^2}{\partial x^2}y^2\frac{\partial^2}{\partial z^2} + x^2\frac{\partial^2}{\partial y^2}\frac{\partial^2}{\partial z^2}\right)u. \tag{4.3.3}$$

The discretization of $\frac{\partial^2}{\partial x^2}u$ by finite differences on an equidistant grid is equivalent to linear system solve with the matrix T_2 (3.4.7). The discretization of the mass term x^2u is equivalent to linear system solve with the matrix D_3 (3.4.1), cf. the discretization of problem (4.1.1). Using tensor product arguments, the structure of (4.3.3), and the fact that the operators $\mathcal{G}_k : C^k(\Omega_3) \mapsto C^k(\Omega_3)$, $\mathcal{G}_k u = y^2 u$ and $\mathcal{F} : C^2(\Omega_3) \mapsto C^0(\Omega_3)$, $\mathcal{F}u = \frac{\partial^2}{\partial x^2}u$ are commute, ($\mathcal{F}\mathcal{G}_2 = \mathcal{G}_0\mathcal{F}$), the

first assertion follows by the definition of the matrix C_7 (3.4.21). In order to prove the second assertion, we rewrite the differential operator of (4.3.2) into the form

$$\begin{aligned} & x^2\left(-2\frac{\partial^2}{\partial y^2}+\frac{1}{y^2}\right)\left(-2\frac{\partial^2}{\partial z^2}+\frac{1}{z^2}\right)u \\ & +\left(-2\frac{\partial^2}{\partial x^2}+\frac{1}{x^2}\right)y^2\left(-2\frac{\partial^2}{\partial z^2}+\frac{1}{z^2}\right)u \\ & +\left(-2\frac{\partial^2}{\partial x^2}+\frac{1}{x^2}\right)\left(-2\frac{\partial^2}{\partial y^2}+\frac{1}{y^2}\right)z^2u. \end{aligned}$$

Using the definition of the matrices T_1 (3.4.2) and C_6 (3.4.20), the lemma has been proved. □

4.3.2 Trilinear elements

Consider (4.3.1) in the weak formulation: Find $u \in \mathbb{H}$ such that

$$a_3(u,v) := \int_{\Omega_3} (\omega(x))^2 u_{yz}v_{yz} + (\omega(y))^2 u_{xz}v_{xz} + (\omega(z))^2 u_{xy}v_{xy} = \int_{\Omega_3} gv$$

holds for all $v \in \mathbb{H}$, where

$$\mathbb{H} = \left\{u \in L^2(\Omega_3), \omega(x)u_{yz}, \omega(y)u_{xz}, \omega(z)u_{xy} \in L^2(\Omega_3), u\,|_{\partial\Omega_3} = 0\right\}$$

with $\Omega_3 = (0,1)^3$, and the weight function $\omega(\xi) = \xi$. We discretize (4.3.1) by trilinear elements and introduce the notation of the subsections 4.1.2 and 4.2.3. The mesh

$$T_k = \bigcup_{i,j,l=0}^{n-1} \mathcal{H}_{i,j,l}$$

is chosen as finite element mesh with the hexahedral elements

$$\mathcal{H}_{i,j,l} = \overline{\tau_i^k \times \tau_j^k \times \tau_l^k}, \tag{4.3.4}$$

where $\tau_i^k = (\frac{i}{n}, \frac{i+1}{n})$. On this mesh, the piecewise trilinear nodal shape functions

$$\phi_{i,j,l}^{(t,k)}(x,y,z) = \phi_i^{(1,k)}(x)\phi_j^{(1,k)}(y)\phi_l^{(1,k)}(z) \quad 1 \le i,j,l \le n-1 \tag{4.3.5}$$

are introduced and the conformal finite element approximation space

$$\mathbb{V}_k^{(t)} = \operatorname{span}\left\{\phi_{i,j,l}^{(t,k)}\right\}_{i,j,l=1}^{n-1}$$

is defined. Then, the Galerkin projection of problem (4.3.1) onto $\mathbb{V}_k^{(t)}$ is:
Find $u^k \in \mathbb{V}_k^{(t)}$ such that

$$a_3(u^k, v^k) = \int_{\Omega_3} gv^k \tag{4.3.6}$$

holds for all $v^k \in \mathbb{V}_k^{(t)}$. Then, (4.3.6) is equivalent to solving the system of linear algebraic finite element equations $K_{t,k}\underline{u} = \underline{g}$, where

$$K_{t,k} = \left[a_3(\phi_I^{(t,k)}, \phi_{I'}^{(t,k)})\right]_{I',I}$$

with the multi-indices $I = (i,j,l)$ and $I' = (i',j',l')$. Using (4.3.5), the left hand side of problem (4.3.6) can be rewritten into the form

$$\begin{aligned} a_3(\phi_I^{(t,k)}, \phi_{I'}^{(t,k)}) \;=\; & \int_{\text{supp }\phi_i^{(1,k)}} x^2 \phi_i^{(1,k)}(x)\phi_{i'}^{(1,k)}(x)\,\mathrm{d}x \\ & \int_{\text{supp }\phi_j^{(1,k)}} (\phi_j^{(1,k)})'(y)(\phi_{j'}^{(1,k)})'(y)\,\mathrm{d}y \int_{\text{supp }\phi_l^{(1,k)}} (\phi_l^{(1,k)})'(z)(\phi_{l'}^{(1,k)})'(z)\,\mathrm{d}z \\ & + \int_{\text{supp }\phi_i^{(1,k)}} (\phi_i^{(1,k)})'(x)(\phi_{i'}^{(1,k)})'(x)\,\mathrm{d}x \\ & \int_{\text{supp }\phi_j^{(1,k)}} y^2 \phi_j^{(1,k)}(y)\phi_{j'}^{(1,k)}(y)\,\mathrm{d}y \int_{\text{supp }\phi_l^{(1,k)}} (\phi_l^{(1,k)})'(z)(\phi_{l'}^{(1,k)})'(z)\,\mathrm{d}z \\ & + \int_{\text{supp }\phi_i^{(1,k)}} (\phi_i^{(1,k)})'(x)(\phi_{i'}^{(1,k)})'(x)\,\mathrm{d}x \\ & \int_{\text{supp }\phi_j^{(1,k)}} (\phi_j^{(1,k)})'(y)(\phi_{j'}^{(1,k)})'(y)\,\mathrm{d}y \int_{\text{supp }\phi_l^{(1,k)}} z^2 \phi_l^{(1,k)}(z)\phi_{l'}^{(1,k)}(z)\,\mathrm{d}z. \end{aligned}$$

Thus, from relations (4.1.6), (4.1.8) and (3.4.22), one concludes

$$K_{t,k} = \frac{2}{3n} C_8.$$

Hence, the following lemma has been proved.

LEMMA 4.8. *The discretization of (4.3.1) by trilinear elements on the mesh (4.3.4) leads to the system of linear algebraic equations $K_{t,k}\underline{u} = \underline{g}$, where $K_{t,k} = \frac{2}{3n} C_8$.*

We have shown in subsection 4.1.2 that

$$T_3 = \frac{1}{4n}\left[2a_s(\phi_i^{(1,k)}, \phi_j^{(1,k)}) + a_{\overline{m}}(\phi_i^{(1,k)}, \phi_j^{(1,k)})\right]_{j,i=1}^{n-1},$$

see relation (4.1.9) and

$$D_5 = 6n^3\left[a_m(\phi_i^{(1,k)}, \phi_j^{(1,k)})\right]_{j,i=1}^{n-1},$$

see relation (4.1.8). Using tensor product arguments, it follows that the discretization of the boundary value problem (4.3.2) by trilinear finite elements on the tensor product mesh (4.3.4) is equivalent to solving the system $C_9\underline{u} = \underline{g}$, see relation (3.4.23) for the definition of C_9, the arguments in the proof of Lemma 4.8 and the definition of the bilinear forms $a_s(\cdot,\cdot)$, $a_m(\cdot,\cdot)$ and $a_{\overline{m}}(\cdot,\cdot)$ in (4.1.3). We summarize these observations in the following remark.

REMARK 4.9. *The discretization of (4.3.2) by trilinear elements on the mesh (4.3.4) leads to the system of linear algebraic equations $\tilde{K}_{t,k}\underline{u} = \underline{g}$, where $\tilde{K}_{t,k} = \frac{8}{3n} C_9$.*

5 Fast solvers for degenerated problems

5.1 Introduction, aim, direct methods

In this chapter, we primarily consider problem (4.2.3): Find $u \in H^1_{0,\omega}(\Omega)$ such that

$$a(u,v) := \int_\Omega (\omega(y))^2 u_x v_x + (\omega(x))^2 u_y v_y = \int_\Omega gv =: \langle g, v\rangle \quad \forall v \in H^1_{0,\omega}(\Omega). \tag{5.1.1}$$

The domain $\Omega = (0,1)^2$ is the unit square. The weight function ω is of the type $\omega(\xi) = \xi$.

REMARK 5.1. *The differential operator in (5.1.1) is not uniformly elliptic in the Sobolev space $H^1_0(\Omega)$, an estimate of the type*

$$a(u,u) \geq \gamma \parallel u \parallel^2_{H^1(\Omega)} \quad \forall u \in H^1_0(\Omega) \tag{5.1.2}$$

with a constant $\gamma > 0$ is not satisfied.

Proof: The piecewise linear hat function $\phi^k_{11} \in \mathbb{V}_k \subset H^1_0(\Omega)$ on level k defined in relation (4.2.4) satisfies

$$\parallel \phi^k_{11} \parallel_{H^1(\Omega)} \geq \mid \phi^k_{11} \mid_{H^1(\Omega)} = 2.$$

By (3.4.18) and (4.2.9), one concludes $a(\phi^k_{11}, \phi^k_{11}) = \frac{14}{6}\frac{1}{n^2}$, where $n = 2^k$. Thus, we have found a sequence $\{\phi^k_{11}\}_{k=1}^\infty$ with $\parallel \phi^k_{11} \parallel^2_{H^1(\Omega)} \geq 4$, but $a(\phi^k_{11}, \phi^k_{11}) \to 0$ for $k \to \infty$. Hence, an estimate of the type (5.1.2) is not possible.□

The integrand on the left hand side in (5.1.1) is of the type $(\nabla u)^T \mathcal{B}(x,y) \nabla v$ with the diffusion tensor

$$\mathcal{B}(x,y) = \begin{bmatrix} y^2 & 0 \\ 0 & x^2 \end{bmatrix}.$$

Therefore, the matrix $\mathcal{B}$ is symmetric and positive definite for all $(x,y) \in \Omega$, but not uniformly positive definite. Moreover, the matrix $\mathcal{B}$ is bounded for each $(x,y) \in \Omega$. Such problems are called degenerated problems. In the past, degenerated problems have been considered relatively rarely. One reason is the unphysical behaviour of the partial differential equation which is quite unusual in technical applications. One work focusing on this type of partial differential equation is the book of Kufner and Sändig [54]. Nowadays, problems of this type become more and more popular because there are stochastic pde's which have a similar structure. An example of a degenerated stochastic partial differential equation is the Black-Scholes partial differential equation which was mentioned in the introduction of this work, cf. equation (1.1).

We consider now the discretization of (5.1.1) by linear elements as described in subsection 4.2.2. As shown in subsection 4.2.2, the Galerkin projection of (5.1.1) onto the space $\mathbb{V}_k$ is equivalent to solving the linear system (4.2.6), namely $K_k \underline{u}_k = \underline{g}_k$ with

$$\begin{aligned} K_k = \left[a(\phi_{lm}^k, \phi_{ij}^k)\right]_{i,j,l,m=1}^{n-1} &= \frac{1}{2n^2} C_4 \\ &= \frac{1}{2n^2} \left(D_4 \otimes T_2 + T_2 \otimes D_4\right). \end{aligned}$$

In this chapter, we will derive fast solution methods for (4.2.6). We are not interested in finding a good finite element mesh in order to approximate u in (5.1.1), only efficient solution methods for the resulting systems $C_4 \underline{u}_k = 2n^2 \underline{g}_k$, or equivalently, $K_k \underline{u}_k = \underline{g}_k$ are focused. Firstly, we note that the matrix C_4 is a sparse matrix with 5-point stencil structure and $\mathcal{O}(n^2)$ nonzero matrix entries, cf. the structure of C_4 in (3.4.18) and the proof of Proposition 3.3.
Therefore, it is important to find a method which solves (4.2.6) in $\mathcal{O}(n^2)$ arithmetical operations. Using the usual Cholesky decomposition with lexicographic ordering of the unknowns, the arithmetical cost is proportional to n^4, and the memory requirement is of order n^3. Using the method of nested dissection developed by George, [32], see subsection 2.2, the arithmetical cost can be reduced to $\mathcal{O}(n^3)$ and the memory requirement to $\mathcal{O}(n^2 \log(1+n))$, if only the nonzero elements of the matrix are stored. However, this method is not arithmetically optimal, too. Moreover, the order of the arithmetical cost and memory requirement cannot be improved by taking another reordering for the Cholesky decomposition, [33].

5.2 Slowly convergent iterative methods

Using iterative methods, no additional memory requirement in order to save the matrix K_k is necessary. However, the speed of convergence of the sequence of iterates $\{\underline{u}^{(m)}\}_{m=1}^{\infty}$ to the exact solution $\underline{u}^*$ depends on the condition number of the matrix K_k. As mentioned in Proposition 3.3, $\kappa(K_k) \geq \mathcal{O}(p^2) = \mathcal{O}(n^2)$. Therefore, efficient preconditioners are needed. For systems of finite element equations arising from the discretization of boundary value problems as e.g. $-u_{xx} - u_{yy} = f$, efficient solution techniques are developed in the last two decades. Examples for such solvers are the preconditioned conjugate gradient (pcg) method with BPX preconditioners, [21], or hierarchical basis preconditioners, [80], and multi-grid methods, [40], [43].
However, the differential operator in (5.1.1) is not spectrally equivalent to the Laplacian. It is an elliptic, but not uniformly elliptic differential operator, cf. (5.1.2). In a certain way, this differential operator can be interpreted as an operator with local anisotropies, where the range of anisotropy ε goes to zero, if the discretization parameter h tends to zero.
A typical anisotropic model problem considered in the literature, see [40], is

$$-\frac{\partial^2 u}{\partial x^2} - \varepsilon \frac{\partial^2 u}{\partial y^2} = f, \quad \varepsilon \text{ small.}$$

One iterative method with a rate of convergence independent of the choice of ε is the multi-grid algorithm with a line Gauß-Seidel (GS) smoother, cf. [43] pp.502–533. Bramble and Zhang,

[22], considered multi-grid methods in a more general case as for the Laplace equation. They proved multi-grid convergence for differential operators of the type $-(f(x,y)u_x)_x-(g(x,y)u_y)_y$, where $0 < g(x,y) \leq g_{max}$ and $0 < f_{min} < f(x,y) < f_{max}$, i.e. one of the coefficients can be arbitrarily small. However, both coefficients can be arbitrarily small in (5.1.1). Thus, we have to find a modified solution technique.

5.3 Multi-grid proof for degenerated problems

In the typical multi-grid proofs, cf. [40], one splits the multi-grid operator in a product of two operators $\mathcal{A}$ and $\mathcal{B}$. One proves a smoothing property, see e.g. [67], [68], for the operator $\mathcal{A}$, whereas an approximation property has to be shown for $\mathcal{B}$. Helpful tools for this aim are the approximation theorems for finite elements as the Aubin-Nitsche-trick. In order to prove such a result, the boundedness and the ellipticity of the bilinear form are required in the Sobolev space $H^1(\Omega)$. However, the ellipticity of the bilinear form (5.1.1) cannot be guaranteed, cf. relation (5.1.2).
Another technique in order to prove a mesh-size independent convergence rate has been introduced by Braess, [15]. In this method, the approximation space $\mathbb{V}_k$ is split into a direct sum of the space $\mathbb{V}_{k-1}$ and a complementary space $\mathbb{W}_k$. One obtains a multiplicative solver for the problem on $\mathbb{V}_k$ by solving the problems on $\mathbb{V}_{k-1}$ and $\mathbb{W}_k$. Schieweck, [70], and Pflaum, [65], have extended this technique. This method does not require regularity assumptions to the bilinear form. Moreover, for triangulations of simple geometry as for (4.2.5), the required assumptions are quite simple to handle.
In this section, we will prove a mesh-size independent convergence rate for a multi-grid algorithm using the ideas of Schieweck and Pflaum. The following remark is important for our aim.

REMARK 5.2. *Note that the bilinear form $a(\cdot,\cdot)$ is positive definite on the space $\mathbb{V}_k$.*

5.3.1 Multi-grid algorithm

The space $\mathbb{V}_k$ is represented as the direct sum

$$\mathbb{V}_k = \mathbb{V}_{k-1} \oplus \mathbb{W}_k,$$

where

$$\mathbb{W}_k = \text{span}\{\phi_{ij}^k\}_{(i,j)\in N_k}, \tag{5.3.1}$$

see e.g. [57], [15], [70], [75], [78]. The index subset $N_k \subset \mathbb{N}^2$ contains the indices of the new nodes on level k and is given by

$$N_k := \{(i,j) \in \mathbb{N}^2, 1 \leq i,j \leq n-1, i = 2m-1 \text{ or } j = 2m-1, m \in \mathbb{N}\}. \tag{5.3.2}$$

Let $u_0 \in \mathbb{V}_k$ be the initial guess. One step $u_1 = MULT(k,u_0,g)$ of the multi-grid algorithm $MULT$ is defined recursively as follows.

ALGORITHM 5.3 ($MULT$). *Set $l = k$.*

- *If $l > 1$, then do*

  1. *Pre-smoothing on $\mathbb{W}_l$: Solve*

  $$a(w, v) = \langle g, v\rangle - a(u_0, v) \quad \forall v \in \mathbb{W}_l$$

  approximately by using ν steps of a simple iterative method S, the approximate solution is $\tilde{w}$. Set $u_0^1 = u_0 + \tilde{w}$.

  2. *Coarse grid correction on $\mathbb{V}_{l-1}$: Find $w \in \mathbb{V}_{l-1}$ such that*

  $$a(w, v) = \langle g, v\rangle - a(u_0^1, v) = \langle r, v\rangle \quad \forall v \in \mathbb{V}_{l-1}.$$

  Compute an approximate solution $\tilde{w}$ by using μ_{l-1} steps of the algorithm $MULT(l-1, 0, r)$. Set $u_0^2 = u_0^1 + \tilde{w}$.

  3. *Post-smoothing on $\mathbb{W}_l$: Solve*

  $$a(w, v) = \langle g, v\rangle - a(u_0^2, v) \quad \forall v \in \mathbb{W}_l$$

  approximately by using ν steps of a simple iterative method S, the approximate solution is $\tilde{w}$. Set $u_1 = u_0^2 + \tilde{w}$.

- *else*

  – *Solve $a(w, v) = \langle g, v\rangle - a(u_0, v) \quad \forall v \in \mathbb{V}_1$ exactly.*

- *end-if.*

REMARK 5.4. *In a standard multi-grid algorithm, the space $\mathbb{W}_l$ in 1. and 3. is replaced by $\mathbb{V}_l$, e.g. the smoother operates on the complete approximation space $\mathbb{V}_l$.*

5.3.2 Algebraic convergence theory for multi-grid

Our aim is to prove the convergence of the multi-grid Algorithm 5.3 $MULT$ in order to solve (4.2.6) using $\mu = \mu_l = 3$ and a special line smoother $S = \underline{S}_{0,k}$ on level k which will be defined in (5.3.49). From [65], [70], the following convergence theorem is known for multi-grid algorithms of the type of the algorithm $MULT$.

THEOREM 5.5. *Let us assume that the following assumptions are fulfilled.*

- *Let $a(\cdot, \cdot)$ be a symmetric and positive definite bilinear form on $\mathbb{V}_k$. Let*

  $$\|\cdot\|_a^2 := a(\cdot, \cdot)$$

  be the energy norm.

- *Let S be a smoother satisfying*

$$\| S^{\nu} w \|_a \leq c\rho^{\nu} \| w \|_a \quad \forall w \in \mathbb{W}_k, \tag{5.3.3}$$

where $0 \leq \rho < 1$ independent of k and $c > 0$.

- *There is a constant $0 \leq \gamma < 1$ independent of k such that*

$$(a(v,w))^2 \leq \gamma^2 a(v,v)a(w,w) \quad \forall w \in \mathbb{W}_k, \forall v \in \mathbb{V}_{k-1} \tag{5.3.4}$$

holds.

- *Let $u_{j+1,k} = MULT(k, u_{j,k}, g)$, let u^* be the exact solution of (4.2.6) and let*

$$\sigma_k = \sup_{u_{j,k} - u^* \in \mathbb{V}_k} \frac{\| u_{j+1,k} - u^* \|_a}{\| u_{j,k} - u^* \|_a} \tag{5.3.5}$$

be the convergence rate of $MULT$ in the energy norm with ν smoothing operations.

Then, the recursion formula

$$\sigma_k \leq \sigma_{k-1}^{\mu_{k-1}} + (1 - \sigma_{k-1}^{\mu_{k-1}})(c\rho^{\nu} + (1 - c\rho^{\nu})\gamma)^2 \tag{5.3.6}$$

is valid.

Proof: This theorem has been proved by Schieweck, Theorem 2.2 of [70] with $\rho = \rho_1 = \rho_3$, and Pflaum, see Theorem 4 of [65]. □

The following lemma of the standard multi-grid theory is helpful for the analysis of the recursion formula (5.3.6).

LEMMA 5.6. *Let $\mu_k = \mu \in \mathbb{N}$, $\mu > 1$, and*

$$\kappa := (c\rho^{\nu} + (1 - c\rho^{\nu})\gamma)^2 < \frac{\mu - 1}{\mu}. \tag{5.3.7}$$

Then, the elements σ_k of the recursion

$$\begin{aligned} \sigma_0 &= 0, \\ \sigma_k &= \sigma_{k-1}^{\mu} + (1 - \sigma_{k-1}^{\mu})\kappa \end{aligned}$$

are contained in the interval $[0, \sigma)$. Furthermore, the equation

$$\sigma = \kappa + \sigma^{\mu}(1 - \kappa)$$

has a solution $\sigma \in (0,1)$. More precisely, the sequence $\{\sigma_k\}_{k=0}^{\infty}$ is monotonic increasing and bounded from above by $\sigma < 1$ for $0 < \kappa < \frac{\mu-1}{\mu}$. Especially, we have

$$\sigma = \lim_{k \to \infty} \sigma_k = \begin{cases} 1 & \text{for} \quad \kappa \geq \frac{1}{2} \\ \frac{\kappa}{1-\kappa} & \text{for} \quad \kappa < \frac{1}{2} \end{cases}$$

for $\mu = 2$ and

$$\sigma = \lim_{k \to \infty} \sigma_k = \begin{cases} 1 & \text{for} \quad \kappa \geq \frac{2}{3} \\ \sqrt{\frac{1}{4} + \frac{\kappa}{1-\kappa}} - \frac{1}{2} & \text{for} \quad \kappa < \frac{2}{3} \end{cases} \tag{5.3.8}$$

for $\mu = 3$.

Proof: The proof can be found in several papers, see e.g. Lemma 3 of [65] or Lemma 3.2 of [70]. □

Using Theorem 5.5 and Lemma 5.6, we can prove a mesh-size independent convergence rate $\sigma < 1$ for a symmetric bilinear form a in the case $\mu = 2$, i.e. the W-cycle, if $\kappa < \frac{1}{2}$.
If $\kappa < \frac{2}{3}$, one can prove a mesh-size independent convergence rate $\sigma < 1$ for $\mu = 3$. The number of smoothing steps which are needed in order to reduce $\kappa < \frac{\mu-1}{\mu} = \frac{2}{3}$ can be determined from (5.3.7). This fact is stated as a remark.

REMARK 5.7. *If $\mu = 3$, $c = 1$ in (5.3.3) and $\gamma^2 < \frac{2}{3}$, $\nu > \frac{\ln(\sqrt{\frac{2}{3}}-\gamma)-\ln(1-\gamma)}{\ln \rho}$ smoothing steps are required.*

5.3.3 Basic definitions and helpful lemmata of the linear algebra

We want to prove a mesh-size independent multi-grid convergence rate for the linear system (4.2.6) via Theorem 5.5. Thus, the bounds for ρ in (5.3.3) and γ^2 in (5.3.4) have to be determined. In a first part, some lemmata are derived which are helpful for this aim. Let us introduce and restate some more notation. By (4.2.4), we have

$$\mathbb{V}_k = \operatorname{span}\{\phi_{ij}^k\}_{i,j=1}^{n-1}.$$

We decompose the space $\mathbb{V}_k$ into the space $\mathbb{V}_{k-1}$ and a space $\mathbb{W}_k$, i.e.

$$\mathbb{V}_k = \mathbb{V}_{k-1} \oplus \mathbb{W}_k,$$

cf. relations (5.3.1) and (5.3.2). In order to prove a sufficient strengthened Cauchy-inequality

$$(a(v,w))^2 \leq \gamma^2 a(v,v)a(w,w) \quad \forall v \in \mathbb{V}_{k-1}, w \in \mathbb{W}_k \tag{5.3.9}$$

with $\gamma^2 < 1$, the bilinear form $a(\cdot,\cdot)$ is split into

$$\begin{aligned} a(v,w) &= \int_\Omega y^2 v_x w_x + x^2 v_y w_y \\ &= \sum_{i,j=0}^{n-1} \int_{\mathcal{E}_{i,j}^k} y^2 v_x w_x + x^2 v_y w_y \\ &= \sum_{i,j=0}^{n-1} a^{\mathcal{E}_{i,j}^k}(v,w). \end{aligned} \tag{5.3.10}$$

DEFINITION 5.8. *Let $\mathbb{V}$ be a space of functions on Ω. Let $\Omega_0 \subset \Omega$. We denote the restriction of $\mathbb{V}$ on Ω_0 by $\mathbb{V}\,|_{\Omega_0}$.*

LEMMA 5.9. *Let $a(\cdot,\cdot)$ be a symmetric, positive definite bilinear form. Under the assumption that*

$$(a^{\mathcal{E}_{i,j}^k}(v,w))^2 \leq \gamma_{\mathcal{E}_{i,j}^k}^2 a^{\mathcal{E}_{i,j}^k}(v,v)a^{\mathcal{E}_{i,j}^k}(w,w) \quad i,j = 0,\ldots,n-1 \tag{5.3.11}$$

for all $v \in \mathbb{V}_k\,|_{\mathcal{E}_{ij}^k}$ *and* $w \in \mathbb{W}_k\,|_{\mathcal{E}_{ij}^k}$*, one has*

$$(a(v,w))^2 \leq \gamma^2 a(v,v)a(w,w) \quad \forall v \in \mathbb{V}_k, w \in \mathbb{W}_k$$

with $\gamma^2 = \max_{i,j} \gamma^2_{\mathcal{E}_{i,j}^k}$.

Proof: The proof is standard, [15], [57]. □
Thus, we can deduce from the local constants $\gamma^2_{\mathcal{E}_{i,j}^k}$ in (5.3.11) to the global one γ^2 in (5.3.9). The following proposition is required for some boundary elements.

PROPOSITION 5.10. *Let* $a(\cdot,\cdot)$ *be any bilinear form. Assume that*

$$(a(v,w))^2 \leq \gamma^2 a(v,v)a(w,w) \quad \forall v \in \mathbb{V},\ \forall w \in \mathbb{W}$$

is valid. Let $\mathbb{V}_0 \subset \mathbb{V}$ *and* $\mathbb{W}_0 \subset \mathbb{W}$*. Then,*

$$(a(v,w))^2 \leq \gamma^2 a(v,v)a(w,w) \quad \forall v \in \mathbb{V}_0,\ \forall w \in \mathbb{W}_0$$

holds.

Proof: The proof is trivial. □
The following lemma, see [39], [75], relates the constant $\gamma^2_{\mathcal{E}_{i,j}^k}$ of the strengthened Cauchy-inequality (5.3.11) to the largest eigenvalue of a generalized eigenvalue problem. In order to formulate it, two definitions are needed.

DEFINITION 5.11. *Let* $a(\cdot,\cdot) : \mathbb{V} \times \mathbb{V} \mapsto \mathbb{R}$ *be any bilinear form. We define*

$$\ker\ a = \{v \in \mathbb{V} : a(v,w) = 0 \quad \forall w \in \mathbb{V}\}$$

as the kernel of the bilinear form a*.*

DEFINITION 5.12. *Let* $\mathbb{X}$ *be a linear (finite dimensional) space,* $\mathbb{Y}$ *a subspace of* $\mathbb{X}$*. We define the difference* $\mathbb{X} \ominus \mathbb{Y}$ *as any linear subspace satisfying*

$$\mathbb{X} = \mathbb{Y} \oplus (\mathbb{X} \ominus \mathbb{Y}).$$

We note that the choice of $\mathbb{X} \ominus \mathbb{Y}$ is not unique.

LEMMA 5.13. *Consider the splitting* $\mathbb{V} \oplus \mathbb{W}$*. Let*

$$\mathbb{V} = \text{span}\{\phi_i\}_{i=1}^n, \quad \mathbb{W} = \text{span}\{\psi_i\}_{i=1}^m,$$

$$G = [a(\phi_i,\phi_j)]_{j,i=1}^n\,,\ H = [a(\phi_i,\psi_j)]_{j,i=1}^{m,n}\,,\ J = [a(\psi_i,\psi_j)]_{j,i=1}^m\,.$$

Furthermore, let

$$\mathbb{V} \cap \mathbb{W} = \{\mathbf{0}\}$$

and

$$\ker a \subset \mathbb{V}.$$

Let us assume that the bilinear form $a(\cdot,\cdot)$ is symmetric and positive semidefinite. Then, the minimal constant γ^2 with

$$a(v,w)^2 \leq \gamma^2 a(v,v)a(w,w) \quad \forall v \in \mathbb{V}, w \in \mathbb{W}$$

is equal to the largest eigenvalue λ of the generalized eigenvalue problem

$$V^T H^T J^{-1} H V \underline{w} = \lambda V^T G V \underline{w}. \tag{5.3.12}$$

The matrix $V \in \mathbb{R}^{n\times q}$, $q \leq n$, is chosen arbitrarily such that $\mathrm{im} V = \mathbb{R}^n \ominus \ker G$ and $\ker V^T = \mathbf{0}$.

Proof: We have

$$a(v,w)^2 \leq \gamma^2 a(v,v)a(w,w) \quad \forall v \in \mathbb{V}, w \in \mathbb{W}, \tag{5.3.13}$$

where γ^2 is as small as possible. For $v \in \ker a$, this inequality is satisfied. Hence, it is equivalent to restrict ourselves to $v \in \mathbb{V} \ominus \ker a$, $v, w \neq 0$. Since a is positive semidefinite, one can write

$$\frac{a(v,w)^2}{a(v,v)a(w,w)} \leq \gamma^2$$

for all $v \in \mathbb{V} \ominus \ker a, w \in \mathbb{W}$ and $v, w \neq 0$. Hence, the inequality (5.3.13) is equivalent to

$$\sup_{\substack{v \in \mathbb{V} \ominus \ker a \\ w \in \mathbb{W} \\ v \neq 0, w \neq 0}} \frac{(a(v,w))^2}{a(v,v)a(w,w)} = \gamma^2. \tag{5.3.14}$$

Now, we transform the left hand side of (5.3.14). Using vectors of $\mathbb{R}^n$, we have

$$\gamma^2 = \sup_{\substack{v \in \mathbb{V} \ominus \ker a \\ w \in \mathbb{W} \\ v \neq 0, w \neq 0}} \frac{(a(v,w))^2}{a(v,v)a(w,w)} = \sup_{\substack{\underline{v} \in \mathbb{R}^n \ominus \ker G \\ \underline{w} \in \mathbb{R}^m \\ \underline{v} \neq \mathbf{0}, \underline{w} \neq \mathbf{0}}} \frac{(\underline{w}^T H \underline{v})^2}{\underline{v}^T G \underline{v}\ \underline{w}^T J \underline{w}}.$$

Because of our assumptions, the matrix J is symmetric and positive definite. Substituting $\underline{u} = J^{\frac{1}{2}} \underline{w}$, one obtains

$$\gamma^2 = \sup_{\substack{\underline{v} \in \mathbb{R}^n \ominus \ker G \\ \underline{u} \in \mathbb{R}^m \\ \underline{u} \neq \mathbf{0}, \underline{v} \neq \mathbf{0}}} \frac{(\underline{u}^T J^{-\frac{1}{2}} H \underline{v})^2}{\underline{v}^T G \underline{v}\ \underline{u}^T \underline{u}}.$$

The right hand side is maximal, if $\underline{u} = J^{-\frac{1}{2}} H \underline{v}$. Inserting this and $\underline{v} = V \underline{y}$, we have

$$\begin{aligned} \gamma^2 &= \sup_{\underline{v} \in \mathbb{R}^n \ominus \ker G, \underline{v} \neq \mathbf{0}} \frac{\underline{v}^T H^T J^{-1} H \underline{v}}{\underline{v}^T G \underline{v}} \\ &= \sup_{\underline{y} \in \mathbb{R}^q, \underline{y} \neq \mathbf{0}} \frac{\underline{y}^T V^T H^T J^{-1} H V \underline{y}}{\underline{y}^T V^T G V \underline{y}}. \end{aligned}$$

This is the largest eigenvalue of the generalized eigenvalue problem

$$V^T H^T J^{-1} H V \underline{y} = \lambda V^T G V \underline{y},$$

i.e. $\lambda_{max} \left((V^T G V)^{-1} V^T H^T J^{-1} H V \right) = \gamma^2$, where $V^T G V$ is symmetric and positive definite. □

The proof of the strengthened Cauchy-inequality relies on an estimate for the eigenvalues of a 2×2 matrix. A useful tool is the next lemma.

LEMMA 5.14. *Let $M \in \mathbb{R}^{2\times 2}$ be a matrix with real eigenvalues and α a real number with*

$$r = 2\alpha - \operatorname{trace}(M) \geq 0 \tag{5.3.15}$$

and

$$q = \det M + \alpha^2 - \alpha \operatorname{trace}(M) \geq 0. \tag{5.3.16}$$

Then, we have

$$\lambda_{max}(M) \leq \alpha.$$

Proof: The characteristical polynomial $p_c(x)$ of a 2×2 matrix M is given by

$$p_c(x) = x^2 - \operatorname{trace}(M) x + \det M. \tag{5.3.17}$$

Set $y = x - \alpha$, then

$$\begin{aligned} p_c(x) &= y^2 + (2\alpha - \operatorname{trace}(M)) y + \det M + \alpha^2 - \alpha \operatorname{trace}(M), \\ &= y^2 + ry + q. \end{aligned} \tag{5.3.18}$$

Because of our assumption, M has real eigenvalues. By (5.3.17) and (5.3.18), this polynomial has 2 real roots. Since (5.3.15) and (5.3.16), both zeros are nonpositive. Hence, the roots $x_{1,2}$ of p_c fulfill $x_{1,2} \leq \alpha$. □

The following lemma, see [4], [79], of the finite element analysis is helpful for the proof of relation (5.3.3). It analyzes the eigenvalue bounds of an assembled matrix by the eigenvalue bounds of the element matrices.

LEMMA 5.15. *Let $\{\mathfrak{K}_i \in \mathbb{R}^{m_i \times m_i}\}_{i=1}^n$ be a finite set of symmetric and positive definite matrices. Let*

$$\mathfrak{K} = \sum_{i=1}^{n} L_i^T \mathfrak{K}_i L_i,$$

where $L_i \in \mathbb{R}^{m_i \times m}$ *and* $\mathfrak{K} \in \mathbb{R}^{m \times m}$. *Furthermore, let* $\mathfrak{C}_i$ *a symmetric and positive definite preconditioner for the matrix* $\mathfrak{K}_i$ *with*

$$\lambda_{min}\left({\mathfrak{C}_i}^{-1}\mathfrak{K}_i\right) = \lambda_i > 0, \quad \lambda_{max}\left({\mathfrak{C}_i}^{-1}\mathfrak{K}_i\right) = \lambda^i > 0, \quad i = 1, \ldots, n. \tag{5.3.19}$$

Let

$$\mathfrak{C} = \sum_{i=1}^{n} L_i^T \mathfrak{C}_i L_i.$$

Then, $\lambda_{min}\left(\mathfrak{C}^{-1}\mathfrak{K}\right) \geq \underline{\lambda}$ *and* $\lambda_{max}\left(\mathfrak{C}^{-1}\mathfrak{K}\right) \leq \overline{\lambda}$ *is valid with*

$$\underline{\lambda} = \min_{i=1,\ldots,n} \lambda_i, \quad \overline{\lambda} = \max_{i=1,\ldots,n} \lambda^i.$$

Proof: For all $\underline{v} \in \mathbb{R}^m$, we can estimate

$$\begin{aligned}
(\mathfrak{K}\underline{v}, \underline{v}) &= \left(\sum_{i=1}^{n} L_i^T \mathfrak{K}_i L_i \underline{v}, \underline{v}\right) \\
&= \sum_{i=1}^{n} (\mathfrak{K}_i L_i \underline{v}, L_i \underline{v}) \\
&\leq \sum_{i=1}^{n} \lambda^i (\mathfrak{C}_i L_i \underline{v}, L_i \underline{v}) \\
&\leq \sum_{i=1}^{n} \overline{\lambda} (\mathfrak{C}_i L_i \underline{v}, L_i \underline{v}) \\
&= \overline{\lambda} (\mathfrak{C}\underline{v}, \underline{v}),
\end{aligned}$$

where it follows $\lambda_{max}\left(\mathfrak{C}^{-1}\mathfrak{K}\right) \leq \overline{\lambda}$. The second assertion can be proved by the same arguments. □

5.3.4 Discussion of the strengthened Cauchy-inequality on subelements $\mathcal{E}_{ij}^k$

Consider the strengthened Cauchy inequality (5.3.4) for the bilinear form

$$a(u, v) = \int_\Omega \omega^2(y) u_x v_x + \omega^2(x) u_y v_y.$$

For $i, j > 0$, we prove the strengthened Cauchy-inequality on $\tau_{ij}^{1,k}$ and $\tau_{ij}^{2,k}$. If $i = 0$, or $j = 0$, the result is shown by proving the strengthened Cauchy-inequality on the macro-elements $\mathcal{E}_{ij}^k$. At first, we determine the stiffness matrix on the macro-elements $\mathcal{E}_{ij}^k$ with respect to the two level basis built by the basis functions of $\mathbb{V}_k\,|_{\mathcal{E}_{ij}^k}$ and $\mathbb{W}_{k+1}\,|_{\mathcal{E}_{ij}^k}$. We start with the introduction of the basis functions on $\mathcal{E}_{ij}^k$. Note that the triangle $\tau_{ij}^{2,k}$ is the union of the triangles $\tau_{2i,2j}^{2,k+1}$, $\tau_{2i+1,2j}^{1,k+1}$,

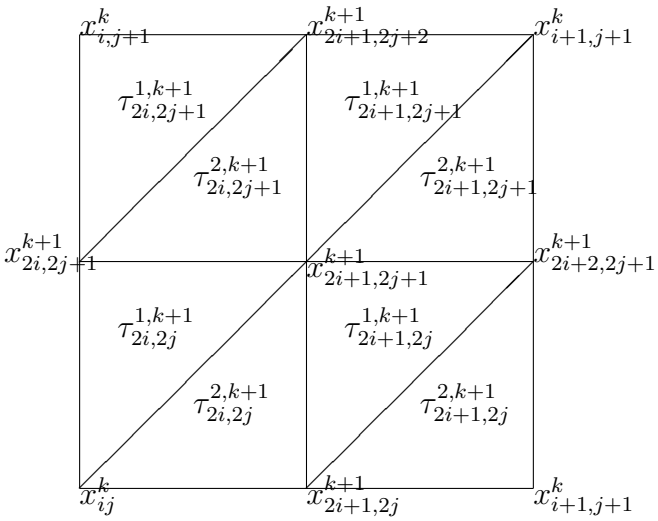

Figure 5.1: Local numbering of the nodes and sub-triangles of $\mathcal{E}^k_{ij}$.

$\tau^{2,k+1}_{2i+1,2j}$, and $\tau^{2,k+1}_{2i+1,2j+1}$, the triangle $\tau^{1,k}_{ij}$ is the union of the triangles $\tau^{1,k+1}_{2i,2j}$, $\tau^{1,k+1}_{2i,2j+1}$, $\tau^{2,k+1}_{2i,2j+1}$, and $\tau^{1,k+1}_{2i+1,2j+1}$. The nodes x^k_{ij}, $x^k_{i,j+1}$, $x^k_{i+1,j}$, and $x^k_{i+1,j+1}$ are the coarse grid nodes, the nodes $x^{k+1}_{2i+1,2j}$, $x^{k+1}_{2i,2j+1}$, $x^{k+1}_{2i+2,2j+1}$, $x^{k+1}_{2i+1,2j+2}$, and $x^{k+1}_{2i+1,2j+1}$ are new in level $k+1$, compare Figure 5.1. Using this notation, we have

$$\mathbb{V}_k \mid_{\mathcal{E}^k_{ij}} = \operatorname{span}\{\phi^k_{lm}\}_{(l,m)\in N^{\mathbb{V}_k}_{ij}} \tag{5.3.20}$$

and

$$\mathbb{W}_{k+1} \mid_{\mathcal{E}^k_{ij}} = \operatorname{span}\{\phi^{k+1}_{lm}\}_{(l,m)\in N^{\mathbb{W}_{k+1}}_{ij}}. \tag{5.3.21}$$

For reasons of simplicity, we write only ϕ^{k+1}_{lm} instead of $\phi^{k+1}_{lm} \mid_{\mathcal{E}^k_{ij}}$ for the restriction of ϕ^{k+1}_{lm} on $\mathcal{E}^k_{ij}$. The index sets in (5.3.20), (5.3.21) are given by

$$\begin{aligned} N^{\mathbb{V}_k}_{ij} &= \{(l,m)\in \mathbb{N}^2_0, i\le l\le i+1, j\le m\le j+1\},\\ N^{\mathbb{W}_{k+1}}_{ij} &= N_{k+1}\cap\{(l,m)\in \mathbb{N}^2_0, 2i\le l\le 2i+2, 2j\le m\le 2j+2\}, \end{aligned}$$

where N_{k+1} was defined in (5.3.2). Because of $\mathbb{V}_k \subset H^1_0(\Omega)$, some modifications are necessary for boundary macro-elements $\mathcal{E}^k_{ij}$, i.e. with $i=0$, $j=0$, $i=n-1$, or $j=n-1$.
On the macro-elements $\mathcal{E}^k_{ij}$, we introduce the matrices

$$\begin{aligned} G_{ij} &:= \left[a^{\mathcal{E}^k_{ij}}(\phi^k_{lm},\phi^k_{rs})\right]_{(r,s),(l,m)\in N^{\mathbb{V}_k}_{ij}},\\ H^T_{ij} &:= \left[a^{\mathcal{E}^k_{ij}}(\phi^{k+1}_{lm},\phi^k_{rs})\right]_{(r,s)\in N^{\mathbb{V}_k}_{ij},(l,m)\in N^{\mathbb{W}_{k+1}}_{ij}},\\ J_{ij} &:= \left[a^{\mathcal{E}^k_{ij}}(\phi^{k+1}_{lm},\phi^{k+1}_{rs})\right]_{(r,s),(l,m)\in N^{\mathbb{W}_{k+1}}_{ij}}. \end{aligned}$$

In the same way, the matrices

$$\begin{aligned} G_{q,ij} &:= \left[a^{\tau_{ij}^{q,k}}(\phi_{lm}^k, \phi_{rs}^k)\right]_{(r,s),(l,m)\in N_{ij}^{q,\mathbb{V}_k}}, \\ H_{q,ij}^T &:= \left[a^{\tau_{ij}^{q,k}}(\phi_{lm}^{k+1}, \phi_{rs}^k)\right]_{(r,s)\in N_{ij}^{q,\mathbb{V}_k},(l,m)\in N_{ij}^{q,\mathbb{W}_{k+1}}}, \\ J_{q,ij} &:= \left[a^{\tau_{ij}^{q,k}}(\phi_{lm}^{k+1}, \phi_{rs}^{k+1})\right]_{(r,s),(l,m)\in N_{ij}^{q,\mathbb{W}_{k+1}}} \end{aligned}$$

with

$$N_{ij}^{q,\mathbb{V}_k} := T_{ij}^q \cap N_{ij}^{\mathbb{V}_k}$$

and

$$N_{ij}^{q,\mathbb{W}_{k+1}} := T_{ij}^q \cap N_{ij}^{\mathbb{W}_{k+1}}$$

are defined on the triangles $\tau_{ij}^{q,k}$, $q = 1, 2$, where $T_{ij}^1 := \{(l,m) \in \mathbb{N}_0^2, l - m \leq i - j\}$ and $T_{ij}^2 := \{(l,m) \in \mathbb{N}_0^2, l - m \geq i - j\}$. The ordering of the rows and columns in the matrices $G_{q,ij}$, $H_{q,ij}$ and $J_{q,ij}$ corresponds to the ordering of the coarse grid nodes and of the new nodes introduced in the beginning of this subsection, cf. Figure 5.1. The entries of the matrices $G_{q,ij}$, $H_{q,ij}$ and $J_{q,ij}$, and G_{ij}, H_{ij} and J_{ij} can be determined by a straightforward calculation. We compute those for the case of a general weight function $\omega(\xi)$ in (5.1.1). The following parameters depending on the integer j are introduced:

$$\begin{aligned} d_j &= \frac{1}{4}\int_{\tau_{2i,2j}^{1,k+1}\cup\tau_{2i,2j+1}^{2,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y), \\ e_j &= \frac{1}{4}\int_{\tau_{2i,2j}^{2,k+1}\cup\tau_{2i+1,2j}^{2,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y), \\ f_j &= \frac{1}{4}\int_{\tau_{2i,2j+1}^{1,k+1}\cup\tau_{2i+1,2j+1}^{1,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y). \end{aligned} \tag{5.3.22}$$

Note that d_j, e_j and f_j are independent of the integer i. The values d_i, e_i and f_i are defined by a permutation of i and j, x and y, and $\tau_{ij}^{2,k}$ and $\tau_{ji}^{1,k}$ in (5.3.22). One obtains the following proposition.

PROPOSITION 5.16. *Let* $0 < i, j < n - 1$. *Then, we have*

$$G_{ij} = \begin{bmatrix} d_i + e_i + d_j + e_j & -d_j - e_j & -d_i - e_i & 0 \\ -d_j - e_j & d_i + f_i + d_j + e_j & 0 & -d_i - f_i \\ -d_i - e_i & 0 & d_i + e_i + d_j + f_j & -d_j - f_j \\ 0 & -d_i - f_i & -d_j - f_j & d_i + f_i + d_j + f_j \end{bmatrix},$$

$$H_{ij}^T = 2\begin{bmatrix} 0 & 0 & -d_j & -d_i & d_i + d_j \\ d_i & 0 & d_j & 0 & -d_i - d_j \\ 0 & d_j & 0 & d_i & -d_i - d_j \\ -d_i & -d_j & 0 & 0 & d_i + d_j \end{bmatrix},$$

$$J_{ij} = 4\begin{bmatrix} d_i+e_j & 0 & 0 & 0 & -d_i \\ 0 & e_i+d_j & 0 & 0 & -d_j \\ 0 & 0 & f_i+d_j & 0 & -d_j \\ 0 & 0 & 0 & d_i+f_j & -d_i \\ -d_i & -d_j & -d_j & -d_i & 2d_i+2d_j \end{bmatrix} \tag{5.3.23}$$

on the macro-elements $\mathcal{E}_{ij}^k$. In the case of matrices on the triangle $\tau_{ij}^{2,k}$, one obtains

$$\begin{aligned} G_{2,ij} &= \begin{bmatrix} d_j+e_j & -d_j-e_j & 0 \\ -d_j-e_j & d_i+f_i+d_j+e_j & -d_i-f_i \\ 0 & -d_i-f_i & d_i+f_i \end{bmatrix}, \\ (H_{2,ij})^T &= 2\begin{bmatrix} 0 & -d_j & d_j \\ d_i & d_j & -d_i-d_j \\ -d_i & 0 & d_i \end{bmatrix}, \\ J_{2,ij} &= 4\begin{bmatrix} d_i+e_j & 0 & -d_i \\ 0 & f_i+d_j & -d_j \\ -d_i & -d_j & d_i+d_j \end{bmatrix}. \end{aligned} \tag{5.3.24}$$

By exchanging the indices i and j in (5.3.24), one obtains the matrices $G_{1,ij} = G_{2,ji}$, $H_{1,ij} = H_{2,ji}$ and $J_{1,ij} = J_{2,ji}$.
In the following, we assume that $\omega(\xi) = \xi$. Thus, there one easily computes

$$\begin{aligned} d_j &= \frac{48j^2+48j+14}{192n^2}, \\ e_j &= \frac{48j^2+16j+2}{192n^2}, \\ f_j &= \frac{48j^2+80j+34}{192n^2}. \end{aligned} \tag{5.3.25}$$

In the case of elements laying on the boundary of the domain Ω, the matrices G_{ij}, H_{ij} and J_{ij} in (5.3.23) and (5.3.24) are similarly defined. However, all rows and columns in G_{ij}, H_{ij} and J_{ij} which correspond to boundary nodes have to be canceled.

COROLLARY 5.17. *We have* $\ker G_{2,ij} \subset \ker H_{2,ij}$ *for* $\tau_{ij}^{2,k}$ *and* $\ker G_{ij} \subset \ker H_{ij}$ *for* $\mathcal{E}_{ij}^k$*, where* $1 \le i,j \le n-2$.

Proof: In the case $\mathcal{E}_{ij}^k$, there one easily derives

$$\ker G_{ij} = \operatorname{span}\{[1,1,1,1]^T\}$$

and for $\tau_{ij}^{2,k}$,

$$\ker G_{2,ij} = \operatorname{span}\{[1,1,1]^T\}.$$

□
Now, we determine the constant $\gamma_{\tau_{ij}^{2,k}}$. For this aim, we prove the next lemma.

LEMMA 5.18. *For $0 < i, j < n-1$, the inequality*

$$(a^{\tau_{ij}^{2,k}}(v,w))^2 \leq \gamma^2_{\tau_{ij}^{2,k}} a^{\tau_{ij}^{2,k}}(v,v) a^{\tau_{ij}^{2,k}}(w,w) \quad \forall v \in \mathbb{V}_k\,|_{\tau_{ij}^{2,k}}, w \in \mathbb{W}_{k+1}\,|_{\tau_{ij}^{2,k}} \tag{5.3.26}$$

holds with $\gamma^2_{\tau_{ij}^{2,k}} = \frac{95}{176}$. The constant is optimal in the case $i = j = 1$.

Proof: Corollary 5.17 states $\ker G_{2,ij} \subset \ker H_{2,ij}$. By Proposition 5.16 and

$$\det J_{2,ij} = d_i e_j f_i + d_i d_j f_i + e_j f_i d_j + d_i e_j d_j > 0$$

(equivalent to $\ker J_{2,ij}$ is trivial), Lemma 5.13 can be applied. We have

$$\ker G_{2,ij} = \operatorname{span}\{[1,1,1]^T\}.$$

Thus, the matrix V can be chosen as

$$V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

The matrix $V^T G_{2,ij} V$ is symmetric and positive definite, and $V^T (H_{2,ij})^T (J_{2,ij})^{-1} H_{2,ij} V$ is symmetric. Therefore, the generalized 2×2 eigenvalue problem

$$V^T (H_{2,ij})^T (J_{2,ij})^{-1} H_{2,ij} V \underline{x} = V^T G_{2,ij} V \lambda \underline{x}$$

has real eigenvalues and is equivalent to the eigenvalue problem

$$M\underline{x} = (V^T G_{2,ij} V)^{-1} V^T (H_{2,ij})^T (J_{2,ij})^{-1} H_{2,ij} V \underline{x} = \lambda \underline{x}.$$

This yields (with a computer algebra system) with $\alpha = \gamma^2_{\tau_{ij}^{2,k}} = \frac{95}{176}$ to

$$r = 2\alpha - \operatorname{trace}(M) \geq 0 \tag{5.3.27}$$

and

$$q = \det M + \alpha^2 - \alpha \operatorname{trace}(M) \geq 0. \tag{5.3.28}$$

Using Lemmata 5.13 and 5.14, we have (5.3.26). □

REMARK 5.19. *We obtain the constant $\gamma^2_{\tau_{ij}^{2,k}} = \frac{95}{176}$ by a direct computation for $i = j = 1$.*

REMARK 5.20. *The values r (5.3.27) and q (5.3.28) are broken rational functions with respect to i and j. We give the exact values in the appendix on page 111.*

Lemma 5.18 has some important corollaries.

COROLLARY 5.21. *For $0 < i, j < n-1$, the inequality*

$$(a^{\tau_{ij}^{1,k}}(v,w))^2 \leq \gamma^2_{\tau_{ij}^{1,k}} a^{\tau_{ij}^{1,k}}(v,v) a^{\tau_{ij}^{1,k}}(w,w) \quad \forall v \in \mathbb{V}_k \mid_{\tau_{ij}^{1,k}}, w \in \mathbb{W}_{k+1} \mid_{\tau_{ij}^{1,k}} \tag{5.3.29}$$

is valid with $\gamma^2_{\tau_{ij}^{1,k}} = \frac{95}{176}$.

Proof: Since the differential operator in (5.1.1) is symmetric with respect to x and y, relation (5.3.26) is valid for the triangle $\tau_{ij}^{1,k}$, too. □

COROLLARY 5.22. *For $0 < i, j < n-1$, the estimate*

$$(a^{\mathcal{E}_{ij}^k}(v,w))^2 \leq \gamma^2_{\mathcal{E}_{ij}^k} a^{\mathcal{E}_{ij}^k}(v,v) a^{\mathcal{E}_{ij}^k}(w,w) \quad \forall v \in \mathbb{V}_k \mid_{\mathcal{E}_{ij}^k}, w \in \mathbb{W}_k \mid_{\mathcal{E}_{ij}^k} \tag{5.3.30}$$

holds with $\gamma^2_{\mathcal{E}_{ij}^k} = \frac{95}{176}$.

Proof: We use the arguments of Lemma 5.9. Then, by Lemma 5.18 and Corollary 5.21, the assertion follows. □
Hence, we have proved a strengthened Cauchy-inequality on the macro-elements $\mathcal{E}_{ij}^k$ for $0 < i, j < n-1$. The remaining cases are the macro-elements $\mathcal{E}_{ij}^k$, where one or both of the indices i or j are equal to 0 or $n-1$. Relatively simple is the case $i = n-1$ or $j = n-1$.

COROLLARY 5.23. *Let $i, j > 0$. The inequality*

$$(a^{\mathcal{E}_{ij}^k}(v,w))^2 \leq \gamma^2_{\mathcal{E}_{ij}^k} a^{\mathcal{E}_{ij}^k}(v,v) a^{\mathcal{E}_{ij}^k}(w,w) \quad \forall v \in \mathbb{V}_k \mid_{\mathcal{E}_{ij}^k}, w \in \mathbb{W}_{k+1} \mid_{\mathcal{E}_{ij}^k} \tag{5.3.31}$$

is valid for $i = n-1$ or $j = n-1$ with $\gamma^2_{\mathcal{E}_{ij}^k} \leq \frac{95}{176}$.

Proof: Consider the case $i = n-1$ and $0 < j < n-1$. We omit the unknowns corresponding to $\phi^k_{i+1,j}$, $\phi^k_{i+1,j+1}$ and $\phi^{k+1}_{2i+2,2j+1}$ in the matrices G_{ij}, H_{ij} and J_{ij} defined in (5.3.23). More precisely, we have to cancel the second and last row and column in G_{ij}, the second row, the fourth row and the third column in H_{ij}^T, and the third row and column in J_{ij}. Note that the assumption $i < n-1$ is not used in the proof of Lemma 5.18. Hence, this estimate and all corollaries of this lemma are valid for $i = n-1$ and $0 < j < n-1$, too. By Lemma 5.10, one can conclude that a Dirichlet boundary condition does not increase the constant of the strengthened Cauchy-inequality. The cases $j = n-1$, $0 < i < n-1$, and $i = j = n-1$ follow by symmetry of the differential operator or with same arguments. □
More difficult is the case $0 < i < n-1$ and $j = 0$. It is not possible to split $\mathcal{E}_{ij}^k$ into $\tau_{ij}^{1,k}$ and $\tau_{ij}^{2,k}$, if $j = 0$ and to prove the strengthened Cauchy-inequalitity on the triangles $\tau_{i,0}^{1,k}$ and $\tau_{i,0}^{2,k}$. On the triangle $\tau_{i,0}^{1,k}$, we have no influence of the Dirichlet boundary condition. We would obtain a constant $\gamma_{\tau_{i,0}^{1,k}}$ which is closer to 1. In order to avoid this phenomenon, $\gamma_{\mathcal{E}_{i,0}^k}$ is estimated directly. The unknowns corresponding to $\phi^k_{i+1,0}$, $\phi^k_{i,0}$ and $\phi^{k+1}_{2i+1,0}$, namely the first two rows and columns of G_{ij} and the first row and column of J_{ij} in (5.3.23), are omitted as the corresponding rows and columns in H_{ij}^T.

By Proposition 5.16,

$$\begin{aligned} G_{i,0} &= \begin{bmatrix} d_i + e_i + 48\eta & -48\eta \\ -48\eta & d_i + f_i + 48\eta \end{bmatrix}, \\ H_{i,0}^T &= 2\begin{bmatrix} 14\eta & 0 & d_i & -d_i - 14\eta \\ -14\eta & 0 & 0 & d_i + 14\eta \end{bmatrix}, \\ J_{i,0} &= 4\begin{bmatrix} e_i + 14\eta & 0 & 0 & -14\eta \\ 0 & f_i + 14\eta & 0 & -14\eta \\ 0 & 0 & d_i + 34\eta & -d_i \\ -14\eta & -14\eta & -d_i & 2d_i + 28\eta \end{bmatrix} \end{aligned}$$

are valid (with $\eta = \frac{1}{192n^2}$). Since $\ker G_{i,0} = \{\mathbf{0}\}$, the identity matrix is a possible choice for V. Using a computer algebra program, we can prove the following lemma with the same arguments as in the proof of Lemma 5.18.

LEMMA 5.24. *The relation*

$$\gamma^2_{\mathcal{E}_{i,0}} < \frac{95}{176} \tag{5.3.32}$$

is valid for $0 < i < n-1$ *and*

$$\gamma^2_{\mathcal{E}_{0,j}} < \frac{95}{176} \tag{5.3.33}$$

is valid for $0 < j < n-1$.

REMARK 5.25. *The estimates (5.3.32) and (5.3.33) can be extended to* $i = n-1$ *and* $j = n-1$ *using the same arguments as in the proof of Corollary 5.23.*

The last case is $i = j = 0$. This case is very simple. By (5.3.25), one has

$$d_0 = 14\eta, \quad e_0 = 2\eta, \quad f_0 = 34\eta$$

with $\eta = \frac{1}{192n^2}$. Furthermore, we note that

$$\mathbb{V}_{k+1}\,|_{\mathcal{E}_{00}^k} = \operatorname{span}\{\phi_{1,1}^k, \phi_{1,0}^{k+1}, \phi_{0,1}^{k+1}, \phi_{0,0}^{k+1}\} \cap H_0^1(\Omega) = \operatorname{span}\{\phi_{1,1}^k\}.$$

From Proposition 5.16, we obtain

$$\begin{aligned} G_{00} &= [2(d_0 + f_0)], \\ H_{00}^T &= \begin{bmatrix} 0 & 0 & 4d_0 \end{bmatrix}, \\ J_{00} &= 4\begin{bmatrix} d_0 + f_0 & 0 & -d_0 \\ 0 & d_0 + f_0 & -d_0 \\ -d_0 & -d_0 & 4d_0 \end{bmatrix} \end{aligned}$$

by canceling the first three rows and columns of G_{ij}, the first two rows and columns of J_{ij}, and the corresponding rows and columns of H_{ij} in (5.3.23). G_{00} is regular. Thus, the matrix $V = [1]$ can be chosen. Using Lemma 5.13, a short computation shows

$$\gamma^2_{\mathcal{E}_{00}} = (V^T G_{00} V)^{-1} V^T H_{00}^T J_{00}^{-1} H_{00} V = \frac{d_0}{d_0 + 2f_0} = \frac{7}{41}. \tag{5.3.34}$$

Now, the main result of this subsection can be stated.

THEOREM 5.26. *Let $\omega(\xi) = \xi$ and let the bilinear form $a(\cdot,\cdot)$ be defined in (5.1.1). Then, the inequality*

$$(a(v,w))^2 \leq \gamma^2 a(v,v)a(w,w) \quad \forall v \in \mathbb{V}_k, w \in \mathbb{W}_{k+1}.$$

is valid with $\gamma^2 = \frac{95}{176}$.

Proof: We apply Lemma 5.9 and estimate $\gamma^2_{\mathcal{E}^k_{ij}}$. The assertion follows by Corollary 5.22 for $0 < i,j < n-1$, by Corollary 5.23 for $0 < i < n-1$ and $j = n-1$, or $0 < j \leq n-1$ and $i = n-1$, by relation (5.3.34) for $i = j = 0$, and by Lemma 5.24 and Remark 5.25 for the remaining cases. □

5.3.5 Construction of the smoother

In order to apply multi-grid to the linear system (4.2.6), we need an efficient smoother. This smoother will be contructed by the local behaviour of the differential operator. An idea of Axelsson and Padiy, [4], for anisotropic problems is extended to bilinear forms as in problem (5.1.1). This smoother operates on the space $\mathbb{W}_{k+1}$ only. We consider the finite element discretization of (4.2.3) with the bilinear form

$$a(u,v) := \int_\Omega (\omega(y))^2 u_x v_x + (\omega(x))^2 u_y v_y \, \mathrm{d}x\mathrm{d}y = \int_\Omega gv \, \mathrm{d}x\mathrm{d}y =: \langle g, v\rangle$$

(see subsection 4.2.2) and a general weight function $(\omega(\xi))^2$.

ASSUMPTION 5.27. *The weight function $(\omega(\xi))^2$ is assumed to be of the form $(\omega(\xi))^2 = \xi^{2\alpha}$ with $\alpha \geq 0$.*

The most interesting case is $\alpha = 1$.
Consider the triangle $\tau^{2,k}_{ij}$. For our discussion, only the sub-matrices $J_{s,ij}$, where $0 \leq i,j \leq n-1$ and $s = 1,2$, are needed which correspond to the nodal basis functions on $\mathbb{W}_{k+1}$. The two cases $i < j$ and $i \geq j$ are discussed. We start with $i < j$. By Proposition 5.16,

$$J_{2,ij} = 4 \begin{bmatrix} d_i + e_j & 0 & -d_i \\ 0 & f_i + d_j & -d_j \\ -d_i & -d_j & d_i + d_j \end{bmatrix}.$$

The index k is omitted. For $i < j$, the matrix

$$\mathfrak{C}_{2,ij} = 4 \begin{bmatrix} d_i + e_j & 0 & 0 \\ 0 & f_i + d_j & -d_j \\ 0 & -d_j & d_i + d_j \end{bmatrix} \tag{5.3.35}$$

is introduced. In the matrix $\mathfrak{C}_{2,ij}$, we set all off diagonal entries of $J_{2,ij}$ to 0 which have relatively small absolute values in comparison to the corresponding main diagonal entries. Since ω is monotonic increasing, the relation $d_i < d_j$ is valid for $i < j$. Thus, we set the $-d_i$ entries of $J_{2,ij}$ in $\mathfrak{C}_{2,ij}$ to 0. We prove now the following lemma.

LEMMA 5.28. *For $0 \leq i < j < n$, the eigenvalue estimates*

$$\begin{aligned} \lambda_{min}\left({\mathbb{C}_{2,ij}}^{-1} J_{2,ij}\right) &\geq 1 - \frac{1}{3}\sqrt{3} \quad \textit{and} \\ \lambda_{max}\left({\mathbb{C}_{2,ij}}^{-1} J_{2,ij}\right) &\leq 1 + \frac{1}{3}\sqrt{3} \end{aligned}$$

hold.

Proof: Let

$$\beta = d_i f_i + d_i d_j + f_i d_j.$$

Then, we have

$$\mathbb{C}_{2,ij}^{-1} J_{2,ij} = \begin{bmatrix} 1 & 0 & \frac{-d_i}{d_i+e_j} \\ \frac{-d_i d_j}{\beta} & 1 & 0 \\ \frac{-d_i f_i - d_i d_j}{\beta} & 0 & 1 \end{bmatrix}.$$

This matrix has the characteristical polynomial

$$\det(\lambda I - \mathbb{C}_{2,ij}^{-1} J_{2,ij}) = (\lambda - 1)\left((1-\lambda)^2 - \frac{d_i}{d_i + e_j}\frac{d_i f_i + d_i d_j}{d_i f_i + d_i d_j + f_i d_j}\right).$$

The roots λ_i, $i = 1, 2, 3$, of this polynomial are

$$\begin{aligned} \lambda_1 &= 1, \\ \lambda_{2,3} &= 1 \pm \sqrt{\rho}, \end{aligned}$$

where

$$\rho = \frac{d_i}{d_i + e_j}\frac{d_i f_i + d_i d_j}{d_i f_i + d_i d_j + f_i d_j}. \tag{5.3.36}$$

Note that for all i and j, the values d_j, e_j and f_j are mean values of the positive function $(\omega(y))^2$ over the union of two triangles having a volume of $\frac{1}{8n^2}$. By the monotony of the weight function, the inequality $d_i \leq f_i$ holds for all $i \in \mathbb{N}$, cf. (5.3.22) and Figure 5.1. Therefore,

$$\frac{d_i f_i + d_i d_j}{d_i f_i + d_i d_j + f_i d_j} \leq \frac{d_i f_i + d_i d_j}{d_i f_i + 2 d_i d_j} = \frac{f_i + d_j}{f_i + 2 d_j} = \frac{1}{1 + \frac{1}{\frac{f_i}{d_j}+1}}.$$

Moreover, by $i \leq j - 1$ and the monotony of the weight function, one has $\omega(x) \leq \omega(y)$ for all $x, y \in \tau_{ij}^{2,k}$. Thus, by integration over sub-triangles of $\tau_{ij}^{2,k}$ with volume $\frac{1}{8n^2}$, cf. Figure 5.1,

$$\begin{aligned} f_i = \frac{1}{4}\int_{\tau_{2i+1,2j}^{2,k+1} \cup \tau_{2i+1,2j+1}^{2,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y) &\leq \frac{1}{4}\int_{\tau_{2i+1,2j}^{1,k+1} \cup \tau_{2i+1,2j+1}^{2,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y) = d_j, \\ d_i = \frac{1}{4}\int_{\tau_{2i,2j}^{2,k+1} \cup \tau_{2i+1,2j}^{1,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y) &\leq \frac{1}{4}\int_{\tau_{2i,2j}^{2,k+1} \cup \tau_{2i+1,2j}^{2,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y) = e_j. \end{aligned}$$

Therefore, we obtain the estimates

$$\frac{d_i f_i + d_i d_j}{d_i f_i + d_i d_j + f_i d_j} \leq \frac{2}{3} \tag{5.3.37}$$

and

$$\frac{d_i}{d_i + e_j} \leq \frac{1}{2}. \tag{5.3.38}$$

Inserting the estimates (5.3.37) and (5.3.38) into (5.3.36), one has

$$1 - \sqrt{\frac{1}{3}} \leq \lambda_3 \leq \lambda_1 \leq \lambda_2 \leq 1 + \sqrt{\frac{1}{3}}.$$

Hence, the assertion follows immediately. □
Now, consider the case $i \geq j$. Introducing the matrix

$$\mathfrak{C}_{2,ij} = 4 \begin{bmatrix} d_i + e_j & 0 & -d_i \\ 0 & f_i + d_j & 0 \\ -d_i & 0 & d_i + d_j \end{bmatrix}, \tag{5.3.39}$$

we will show that $\kappa\left(\mathfrak{C}_{2,ij}{}^{-1} J_{2,ij}\right) \leq c$ independent of the parameters j, i, and n. In order to prove this result, the following estimate concerning the weight function is necessary.

LEMMA 5.29. *Let $\omega(\cdot)$ satisfy Assumption 5.27. Then, one has the inequality*

$$0 \leq \left(\omega\left(y + \frac{1}{2n}\right)\right)^2 \leq c(\omega(y))^2 \quad \forall y \geq \frac{1}{n}, \tag{5.3.40}$$

where the constant c is independent of n and y.

The inequality

$$\left(\frac{\xi + 1.5}{\xi + 1}\right)^{2\alpha} = \left(1 + \frac{1}{2\xi + 2}\right)^{2\alpha} \leq \left(\frac{3}{2}\right)^{2\alpha} = c$$

holds for all $\xi \geq 0$ and $\alpha \geq 0$ with $c = \left(\frac{3}{2}\right)^{2\alpha}$. Thus,

$$\begin{aligned} \left(\xi + \frac{3}{2}\right)^{2\alpha} &\leq c\,(\xi + 1)^{2\alpha}, \quad \text{or} \\ \left(\frac{\xi + \frac{3}{2}}{n}\right)^{2\alpha} &\leq c\left(\frac{\xi + 1}{n}\right)^{2\alpha} \end{aligned}$$

with some $n > 0$. Using $(\omega(\xi))^2 = \xi^{2\alpha}$, we have $\left(\omega\left(\frac{\xi+\frac{3}{2}}{n}\right)\right)^2 \leq c\left(\omega\left(\frac{\xi+1}{n}\right)\right)^2$, or, substituting $y = \frac{\xi+1}{n}$,

$$0 \leq \left(\omega\left(y + \frac{1}{2n}\right)\right)^2 \leq c(\omega(y))^2 \quad \forall y \geq \frac{1}{n}$$

which is the desired result. □

LEMMA 5.30. *For $0 \leq j \leq i < n$, one has*

$$\begin{aligned} \lambda_{min}\left(\mathfrak{C}_{2,ij}{}^{-1} J_{2,ij}\right) &\asymp 1 \quad \textit{and} \\ \lambda_{max}\left(\mathfrak{C}_{2,ij}{}^{-1} J_{2,ij}\right) &\asymp 1. \end{aligned}$$

The constants are independent of i, j and n. For $\omega(\xi) = \xi$, the eigenvalue estimates

$$\begin{aligned} \lambda_{min}\left(\mathfrak{C}_{2,ij}{}^{-1} J_{2,ij}\right) &\geq 1 - \frac{2}{11}\sqrt{11} \quad \textit{and} \\ \lambda_{max}\left(\mathfrak{C}_{2,ij}{}^{-1} J_{2,ij}\right) &\leq 1 + \frac{2}{11}\sqrt{11} \end{aligned}$$

are valid.

Proof: We start with the case $i < n-1$ and $j > 0$. The proof is similar to the proof of Lemma 5.28. A short calculation yields

$$\det(\lambda I - \mathfrak{C}_{2,ij}^{-1} J_{2,ij}) = (\lambda - 1)\left((\lambda-1)^2 - \frac{d_j}{d_j + f_i}\frac{d_i d_j + e_j d_j}{d_i e_j + d_i d_j + e_j d_j}\right).$$

By $i \geq j$ and the monotony of the weight function ω, we have

$$\begin{aligned} \int_{\tau_{2i+1,2j}^{2,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y) &= \int_{\tau_{2i+1,2j+1}^{2,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y) \\ &\geq \int_{\tau_{2j+1,2i}^{2,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y) \\ &= \int_{\tau_{2i,2j+1}^{1,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y) \\ &\geq \int_{\tau_{2i+1,2j}^{1,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y). \end{aligned} \tag{5.3.41}$$

For the same reason,

$$\int_{\tau_{2i,2j}^{2,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y) \leq \int_{\tau_{2i+1,2j}^{1,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y). \tag{5.3.42}$$

Using (5.3.41) and (5.3.42),

$$f_i = \int_{\tau_{2i+1,2j}^{2,k+1} \cup \tau_{2i+1,2j+1}^{2,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y) \geq \int_{\tau_{2i,2j}^{2,k+1} \cup \tau_{2i+1,2j}^{1,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y) = d_j. \tag{5.3.43}$$

By Lemma 5.29, we have

$$0 \leq \left(\omega\left(y + \frac{1}{2n}\right)\right)^2 \leq c(\omega(y))^2 \quad \forall y \geq \frac{1}{n}.$$

Integration over $\tau^{2,k+1}_{2i+1,2j}$ gives

$$\int_{\tau^{2,k+1}_{2i+1,2j}} \left(\omega\left(y+\frac{1}{2n}\right)\right)^2 \,\mathrm{d}(x,y) \le c \int_{\tau^{2,k+1}_{2i+1,2j}} (\omega(y))^2 \,\mathrm{d}(x,y)$$

with $j \ge 1$. With a change of variables $\tilde{y} = y + \frac{1}{2n}$ in the left integral, the integration will be done now over $\tau^{2,k+1}_{2i+1,2j+1}$,

$$\int_{\tau^{2,k+1}_{2i+1,2j+1}} (\omega(y))^2 \,\mathrm{d}(x,y) \le c \int_{\tau^{2,k+1}_{2i+1,2j}} (\omega(y))^2 \,\mathrm{d}(x,y). \tag{5.3.44}$$

Using (5.3.44),

$$\int_{\tau^{2,k+1}_{2i,2j}} (\omega(y))^2 \,\mathrm{d}(x,y) = \int_{\tau^{2,k+1}_{2i+1,2j}} (\omega(y))^2 \,\mathrm{d}(x,y),$$

and

$$\int_{\tau^{1,k+1}_{2i+1,2j}} (\omega(y))^2 \,\mathrm{d}(x,y) \le \int_{\tau^{2,k+1}_{2i+1,2j+1}} (\omega(y))^2 \,\mathrm{d}(x,y),$$

we have

$$d_j = \frac{1}{4} \int_{\tau^{1,k+1}_{2i+1,2j} \cup \tau^{2,k+1}_{2i+1,2j+1}} (\omega(y))^2 \,\mathrm{d}(x,y) \le \frac{c}{4} \int_{\tau^{2,k+1}_{2i,2j} \cup \tau^{2,k+1}_{2i+1,2j}} (\omega(y))^2 \,\mathrm{d}(x,y) = e_j. \tag{5.3.45}$$

For the case $\alpha = 1$, the constant c can be chosen by the more accurate estimate $c = \frac{5}{3}$, cf. the explicit structure of d_j and e_j in (5.3.25). Using (5.3.45) and $d_j \le d_i$ for $j \le i$, one can estimate

$$e_j d_j + d_j d_i \le (c+1) e_j d_i.$$

Equivalently, one obtains

$$(c+2)(e_j d_j + d_j d_i) \le (c+1)(e_j d_i + e_j d_j + d_j d_i).$$

Together with (5.3.43), the assertion follows as in the proof of Lemma 5.28.
Consider now $i = n-1$. Then, the second row and column of $\mathfrak{C}_{2,ij}$ and $J_{2,ij}$ has to be canceled. Thus, the matrices $\mathfrak{C}_{2,n-1,j}$ and $J_{2,n-1,j}$ are identical and

$$\lambda_1(\mathfrak{C}^{-1}_{2,n-1,j} J_{2,n-1,j}) = \lambda_2(\mathfrak{C}^{-1}_{2,n-1,j} J_{2,n-1,j}) = 1.$$

The last case is $j = 0$. We have to omit the first row and column in $\mathfrak{C}_{2,i,0}$ and $J_{2,i,0}$. A short calculation shows

$$\det(\lambda I - \mathfrak{C}^{-1}_{2,i,0} J_{2,i,0}) = (\lambda-1)^2 - \frac{d_0}{f_i + d_0} \frac{d_0}{d_0 + d_i}.$$

Since $d_0 \le d_i$ and $d_0 \le f_i$ for $i \ge 0$, cf. relation (5.3.43), $\frac{d_0}{d_0+d_i} \le \frac{1}{2}$ and $\frac{d_0}{d_0+f_i} \le \frac{1}{2}$ follows. Hence, the estimates

$$\frac{1}{2} \le \lambda_2 < \lambda_1 \le \frac{3}{2}$$

are obtained for the roots of the characteristical polynomial of the matrix $\mathbb{C}_{2,i,0}^{-1} J_{2,i,0}$. □
In (5.3.35), (5.3.39), we have defined a local preconditioner $\mathbb{C}_{2,ij}$ for the macro-element stiffness matrices $J_{2,ij}$ corresponding to the triangle $\tau_{ij}^{2,k}$. On the triangles $\tau_{ij}^{1,k}$, we define matrices $\mathbb{C}_{1,ij}$ in the same way as $\mathbb{C}_{2,ij}$ for $\tau_{ij}^{2,k}$:

$$\mathbb{C}_{1,ij} = \begin{cases} 4 \begin{bmatrix} e_i + d_j & 0 & -d_j \\ 0 & d_i + f_j & 0 \\ -d_j & 0 & d_i + d_j \end{bmatrix} & \text{for} \quad i \leq j, \\ 4 \begin{bmatrix} e_i + d_j & 0 & 0 \\ 0 & d_i + f_j & -d_i \\ 0 & -d_i & d_i + d_j \end{bmatrix} & \text{for} \quad i > j. \end{cases} \tag{5.3.46}$$

REMARK 5.31. *By the symmetry of the differential operator with respect to the variables x and y, we obtain the same results for the triangles $\tau_{ij}^{1,k}$ as in Lemmata 5.28 and 5.30.*

Now, a global preconditioner $\mathbb{C}_{\mathbb{W}_{k+1}}$ for $K_{\mathbb{W}_{k+1}}$ is defined using the local matrices $\mathbb{C}_{s,ij}$, where $0 \leq i, j \leq n-1$, $s = 1, 2$. The matrix $K_{\mathbb{W}_{k+1}}$ is defined as stiffness matrix K_{k+1} (4.2.6) restricted to the space $\mathbb{W}_{k+1}$, i.e.

$$K_{\mathbb{W}_{k+1}} = \left[a(\phi_{lm}^{k+1}, \phi_{ij}^{k+1}) \right]_{(i,j),(l,m) \in N_{k+1}}$$

(compare (5.3.2), (5.3.3)). The matrix $K_{\mathbb{W}_{k+1}}$ is the result of assembling the local stiffness matrices $J_{s,ij}$, $s = 1, 2$ and $i, j = 0, \ldots, n-1$, i.e.

$$K_{\mathbb{W}_{k+1}} = \sum_{s=1}^{2} \sum_{i,j=0}^{n-1} L_{s,ij}^T J_{s,ij} L_{s,ij}. \tag{5.3.47}$$

The matrices $L_{s,ij} \in \mathbb{R}^{3 \times 3 \cdot 4^{k-1} - 2^k}$ are the usual finite element connectivity matrices. Since

$$(2^k - 1)^2 - (2^{k-1} - 1)^2 = 3 \cdot 4^{k-1} - 2^k,$$

the dimension of $L_{s,ij}$ is $3 \times 3 \cdot 4^{k-1} - 2^k$.

DEFINITION 5.32. *We define the matrix $\mathbb{C}_{\mathbb{W}_{k+1}}$ by*

$$\mathbb{C}_{\mathbb{W}_{k+1}} = \sum_{s=1}^{2} \sum_{i,j=0}^{n-1} L_{s,ij}^T \mathbb{C}_{s,ij} L_{s,ij}. \tag{5.3.48}$$

Because of the properties of the local preconditioners $\mathbb{C}_{s,ij}$, the matrix $\mathbb{C}_{\mathbb{W}_{k+1}}$ is a good preconditioner for $K_{\mathbb{W}_{k+1}}$. This result is stated as the main theorem of this subsection.

THEOREM 5.33. *Let $\omega(\xi)$ satisfy Assumption 5.27, let $\mathbb{C}_{\mathbb{W}_{k+1}}$ and $K_{\mathbb{W}_{k+1}}$ be defined in (5.3.48) and (5.3.47), respectively. Then, one obtains*

$$\begin{aligned} \lambda_{min} \left((\mathbb{C}_{\mathbb{W}_{k+1}})^{-1} K_{\mathbb{W}_{k+1}} \right) &\asymp 1, \\ \lambda_{max} \left((\mathbb{C}_{\mathbb{W}_{k+1}})^{-1} K_{\mathbb{W}_{k+1}} \right) &\asymp 1. \end{aligned}$$

In the case $\omega(\xi) = \xi$, the eigenvalue estimates

$$\begin{aligned} \lambda_{min}\left((\mathfrak{C}_{\mathbb{W}_{k+1}})^{-1}K_{\mathbb{W}_{k+1}}\right) &\geq 1 - \frac{2}{11}\sqrt{11}, \\ \lambda_{max}\left((\mathfrak{C}_{\mathbb{W}_{k+1}})^{-1}K_{\mathbb{W}_{k+1}}\right) &\leq 1 + \frac{2}{11}\sqrt{11} \end{aligned}$$

are valid.

Proof: By (5.3.47) and (5.3.48), the assumptions of Lemma 5.15 are satisfied for the matrices $J_{s,ij}$ and $\mathfrak{C}_{s,ij}$. By Lemma 5.28 and Lemma 5.30, and Remark 5.31, the assertions follow. □

REMARK 5.34. *This result can be extended to more general weight functions ω. The weight function should fulfill an estimate of the type (5.3.45) which means that the weight function does not change rapidly. Another possible assumption is that the weight function $\omega(\xi) \geq 0$ satisfies the following properties:*

- *ω is monotonic increasing,*
- *ω is Lipschitz-continuous with a Lipschitz constant L,*
- *$\omega(\xi) \geq \frac{c}{\xi}$ for $\xi \in (0, \delta)$, $\delta > 0$ with some $c > 0$.*

Proof: Using the last assumption and the monotony of ω,

$$\omega(y) \geq \frac{c}{2n} \quad \forall y \geq \frac{1}{n}.$$

Therefore, $\frac{L}{2n} + \omega(y) \leq \left(1 + \frac{L}{c}\right)\omega(y)$. By the monotony of w and the Lipschitz continuity, one derives

$$\omega\left(y + \frac{1}{2n}\right) \leq \frac{L}{2n} + \omega(y) \leq \left(1 + \frac{L}{c}\right)\omega(y)$$

which gives (5.3.45).□
Applying Theorem 5.33, a preconditioned Richardson iteration can be built as a preconditioned simple iteration method. The error transion operator $\underline{S}_{0,k+1}$ of this method is defined by

$$S_{0,k+1} = I - \zeta(\mathfrak{C}_{\mathbb{W}_{k+1}})^{-1}K_{\mathbb{W}_{k+1}}, \tag{5.3.49}$$

where $S_{0,k+1}$ denotes the matrix representation of $\underline{S}_{0,k+1}$ by the usual fem-isomorphism. This smoother $S = \underline{S}_{0,k+1}$ can be used for the Algorithm $MULT$.

COROLLARY 5.35. *Let*

$$\| w \|_a^2 = a(w, w)$$

be the energy norm of the bilinear form a. Then, for all $w \in \mathbb{W}_{k+1}$

$$\| \underline{S}_{0,k+1}^{\nu} w \|_a \leq \rho_k^{\nu} \| w \|_a$$

holds, where

$$\zeta = 1$$

is the optimal choice of ζ *and* $\rho_k \leq \rho < 1$. *Especially,*

$$\rho = \frac{2}{11}\sqrt{11} \tag{5.3.50}$$

holds for $\omega(\xi) = \xi$.

Proof: By calculation and the definition of the smoother in (5.3.49), we have

$$\begin{aligned}
\rho^2 &= \sup_{w \in \mathbb{W}_{k+1}, w \neq \mathbf{0}} \frac{\| \underline{S}_{0,k+1} w \|_a^2}{\| w \|_a^2} \\
&= \sup_{\underline{w}} \frac{(K_{\mathbb{W}_{k+1}} S_{0,k+1}\underline{w}, S_{0,k+1}\underline{w})}{(K_{\mathbb{W}_{k+1}}\underline{w}, \underline{w})} \\
&= \sup_{\underline{u}} \frac{((K_{\mathbb{W}_{k+1}})^{-\frac{1}{2}} S_{0,k+1}^T K_{\mathbb{W}_{k+1}} S_{0,k+1}(K_{\mathbb{W}_{k+1}})^{-\frac{1}{2}}\underline{u}, \underline{u})}{(\underline{u}, \underline{u})} \\
&= \lambda_{max}((K_{\mathbb{W}_{k+1}})^{-\frac{1}{2}} S_{0,k+1}^T K_{\mathbb{W}_{k+1}} S_{0,k+1}(K_{\mathbb{W}_{k+1}})^{-\frac{1}{2}}) \\
&= \lambda_{max}\left((I - \zeta K_{\mathbb{W}_{k+1}}{}^{\frac{1}{2}}(\mathfrak{C}_{\mathbb{W}_{k+1}})^{-1} K_{\mathbb{W}_{k+1}}{}^{\frac{1}{2}})^2\right) \\
&= \left(\lambda_{max}(I - \zeta K_{\mathbb{W}_{k+1}}{}^{\frac{1}{2}} \mathfrak{C}_{\mathbb{W}_{k+1}}{}^{-1} K_{\mathbb{W}_{k+1}}{}^{\frac{1}{2}})\right)^2 \\
&= (\lambda_{max}(S_{0,k+1}))^2.
\end{aligned}$$

The assertion follows using Theorem 5.33.□

5.3.6 Application of the multi-grid theory to $-x^2u_{yy} - y^2u_{xx} = g$

We apply now the theory of subsection 5.3.2 to problem (4.2.5) with the weight function $\omega(\xi) = \xi$. By Theorem 5.26, assumption (5.3.4) is fulfilled with $\gamma^2 \leq \frac{95}{176}$. The second assumption, (5.3.3), of Theorem 5.5 is fulfilled for the smoother $\underline{S}_{0,k}$ defined in (5.3.49), cf. Corollary 5.35. Hence, we can prove a bound $\sigma < 1$ for the convergence rate of the multi-grid Algorithm 5.3 $MULT$ for $\mu \geq 3$, if we do sufficiently many smoothing steps. The convergence rate $\sigma < 1$ does not depend on the level number k. Since $\gamma^2 > \frac{1}{2}$, no mesh-size independent convergence rate can be proved for $\mu \leq 2$. We summarize the results in the following theorem.

THEOREM 5.36. *Consider the linear system (4.2.6) with the exact solution* u^*. *For* $j = 1, \ldots$, *let the new iterate* $u_{j,k}$ *be defined recursively as* $u_{j+1,k} = MULT(k, u_{j,k}, g)$. *Let us assume that* $\mu = \mu_l \geq 3$ *for* $l = 1, \ldots, k$ *and* $\nu \geq 3$. *Then, the rate of convergence*

$$\sigma_k = \sup_{u_{j,k} - u^* \in \mathbb{V}_k} \frac{\| u_{j+1,k} - u^* \|_a}{\| u_{j,k} - u^* \|_a}$$

on level k *can be bounded by*

$$\sigma_k \leq \sigma < 1.$$

ν	σ
< 2	1
3	0.89385
4	0.80549
8	0.70649
∞	0.69283

Table 5.1: Estimates for the bounds σ of the convergence rates σ_k for $\mu = 3$.

Proof: If $\kappa < \frac{2}{3}$, the assertion follows by Theorem 5.5, cf. relation (5.3.7). Using Lemma 5.6, the number of smoothing steps ν required for a convergence rate $\sigma < 1$ can be analyzed. We have

$$\kappa = c\rho^\nu + (1 - c\rho^\nu)\gamma^2$$

with $c = 1$, $\gamma^2 = \frac{95}{176}$ and $\rho = \frac{2}{11}\sqrt{11}$ by relation (5.3.50). Using Remark 5.7, we have a mesh-size independent convergence rate $\sigma_k \leq \sigma < 1$ for

$$\nu \approx 2.33,$$

i.e. $\nu \geq 3$. □
Table 5.1 displays the bounds of the theoretical convergence rates σ_k for several values of ν obtained by Lemma 5.6 for $\mu = 3$.

5.4 AMLI method

In section 5.3, a mesh-size independent convergence rate has been proved for the Algorithm 5.3. The two main ingredients in this proof are the estimate for the constant of the strengthened Cauchy-inequality (5.3.4) and the construction of a smoother $S_{0,k} = I - (\mathfrak{C}_{\mathbb{W}_k})^{-1} K_{\mathbb{W}_k} \leftrightarrow \underline{S}_{0,k}$ which satisfies (5.3.3). Equivalent to relation (5.3.3) is, cf. Corollary 5.35, that $\kappa\left((\mathfrak{C}_{\mathbb{W}_k})^{-1} K_{\mathbb{W}_k}\right)$ is bounded by some constant c independent of the mesh-size h.
Another multi-level method is the preconditioned conjugate gradient method (pcg) with Algebraic Multi-Level Iteration preconditioner (AMLI) $\tilde{C}_{k,r,\mu}$ derived by Axelsson and Vassilevski, [5], [6]. Let K_k be the stiffness matrix (4.2.6) for the discretization of problem (4.2.3) on page 32. One can show that $\kappa\left(\tilde{C}^{-1}_{k,r,\mu} K_k\right)$ is bounded by some constant c for all $k \in \mathbb{N}$ under the following assumptions:

- the relation (5.3.4) is valid with some constant $\gamma^2 < 1$,
- $\kappa\left((\mathfrak{C}_{\mathbb{W}_k})^{-1} K_{\mathbb{W}_k}\right) < c$ is valid with a constant c independent of the mesh-size h.

In subsection 5.4.1, we will give a general definition of the AMLI preconditioner. In subsection 5.4.2, we will introduce a special AMLI preconditioner $\tilde{C}_{k,r,\mu}$ for K_k (4.2.6) and will show that $\kappa\left(\tilde{C}^{-1}_{k,r,\mu} K_k\right) < c$ for all $k \in \mathbb{N}$.

5.4.1 Convergence theory for AMLI

We define now the Algebraic Multi-Level Iteration preconditioner (AMLI) of Axelsson and Vassilevski, [5], [6]. Consider the stiffness matrix K_k (4.2.6). We assume that the unknowns are ordered in such a way that

$$K_k = \begin{bmatrix} K_{11,k} & K_{12,k} \\ K_{21,k} & K_{22,k} \end{bmatrix},$$

where $K_{22,k} = K_{\mathbb{W}_k}$ corresponds to the nodal basis functions in $\mathbb{W}_k$ and

$$K_{11,k} = \left[a(\phi^k_{2l,2m}, \phi^k_{2i,2j})\right]^{\frac{n}{2}-1}_{i,j,l,m=1}$$

corresponds to nodal basis functions of nodes on level $k-1$. Let $\tilde{C}_{22,l}$ be a preconditioner for $K_{22,l}$ satisfying

$$\begin{aligned} \lambda_{min}\left({K_{22,l}}^{-1}\tilde{C}_{22,l}\right) &\geq 1, \\ \lambda_{max}\left({K_{22,l}}^{-1}\tilde{C}_{22,l}\right) &\leq 1+b \end{aligned} \tag{5.4.1}$$

with some constant $b \geq 0$ for $l = 1, \ldots, k$. We introduce

$$\hat{K}_k = \begin{bmatrix} \hat{K}_{11,k} & \hat{K}_{12,k} \\ \hat{K}_{21,k} & K_{22,k} \end{bmatrix} \tag{5.4.2}$$

which is the stiffness matrix with respect to the two level basis, i.e.

$$\{\phi^{k-1}_{ij}\}^{\frac{n}{2}-1}_{i,j=1} \in \mathbb{V}_{k-1}$$

and

$$\{\phi^k_{ij}\},\ \phi^k_{ij} \in \mathbb{W}_k.$$

This basis corresponds to the splitting $\mathbb{V}_k = \mathbb{V}_{k-1} \oplus \mathbb{W}_k$. We assume that there exists a constant $\gamma^2 < 1$, the constant in the strengthened Cauchy-inequality, with

$$\gamma^2 = \sup_{\substack{v \in \mathbb{V}_{k-1} \\ w \in \mathbb{W}_k \\ v \neq 0, w \neq 0}} \frac{(a(v,w))^2}{a(v,v) \cdot a(w,w)}, \tag{5.4.3}$$

or, equivalently,

$$(a(v,w))^2 \leq \gamma^2 a(v,v) \cdot a(w,w) \quad \forall v \in \mathbb{V}_{k-1}, w \in \mathbb{W}_k.$$

From (5.4.2), we have

$$\hat{K}_{11,k} = K_{k-1}.$$

Obviously, there

$$\hat{K}_k = \mathfrak{I}_k K_k \mathfrak{I}_k^T$$

holds with the finite element interpolation matrix

$$\mathfrak{I}_k = \begin{bmatrix} I & \mathfrak{I}_{12,k} \\ \mathbf{0} & I \end{bmatrix}.$$

We define now, see [6], [50], the AMLI preconditioning matrix $\tilde{C}_{k,r,\mu}$.

DEFINITION 5.37. *Let $P_{\mu,r}$ be a polynomial of degree μ satisfying*

$$P_{\mu,r}(0) = 1 \tag{5.4.4}$$

and

$$0 < P_{\mu,r}(t) < 1 \quad \textit{for} \quad 0 < t \leq 1$$

and $r \in \mathbb{R}$. Let $\tilde{C}_{22,k}$ be a matrix which fulfills (5.4.1). Then, we define the preconditioning matrix $\tilde{C}_{k,r,\mu}$ recursively by

$$\tilde{C}_{k,r,\mu} = \begin{cases} \begin{bmatrix} \tilde{C}^c_{k-1,r,\mu} & K_{12,k} + \mathfrak{I}_{12,k}(K_{22,k} - \tilde{C}_{22,k}) \\ \mathbf{0} & \tilde{C}_{22,k} \end{bmatrix} & \\ \times \begin{bmatrix} I & \mathbf{0} \\ \tilde{C}^{-1}_{22,k}(K_{21,k} + (K_{22,k} - \tilde{C}_{22,k})\mathfrak{I}^T_{12,k}) & I \end{bmatrix} & \textit{for} \quad k \geq 2, \\ K_k & \textit{for} \quad k = 1 \end{cases} \tag{5.4.5}$$

with

$$(\tilde{C}^c_{k-1,r,\mu})^{-1} = (I - P_{\mu,r}(\tilde{C}^{-1}_{k-1,r,\mu} K_{k-1}))K^{-1}_{k-1}. \tag{5.4.6}$$

Examples for the choice of the polynomial $P_{\mu,r}$ are given in [5], [6]. We consider there

$$P_{\mu,\frac{2}{1+\alpha}}(t) = \frac{T_\mu\left(\frac{1+\alpha-2t}{1-\alpha}\right) + 1}{T_\mu\left(\frac{1+\alpha}{1-\alpha}\right) + 1} \tag{5.4.7}$$

with some $0 < \alpha < 1$ ($r = \frac{2}{1+\alpha}$), where $T_\mu(x)$ denotes the μ-th Chebyshev-polynomial first kind, i.e.

$$T_\mu(x) = \cos(\mu \arccos(x)).$$

The following theorem is valid.

THEOREM 5.38. *Consider the preconditioner $\tilde{C}_{k,r,\mu}$ (5.4.5) with the polynomial defined by relation (5.4.7). Let us assume that*

$$\mu > \frac{1}{\sqrt{1-\gamma^2}}. \tag{5.4.8}$$

Thus, the two eigenvalue estimates $\lambda_{min}\left(\tilde{C}_{k,r,\mu}^{-1}K_k\right) \geq c_{17}$ *and* $\lambda_{max}\left(\tilde{C}_{k,r,\mu}^{-1}K_k\right) \leq 1$ *hold for all* $k \in \mathbb{N}$*, where*

$$c_{17} = (1-\gamma^2)\left(b + \left(\frac{(1+\sqrt{\alpha})^\mu + (1-\sqrt{\alpha})^\mu}{(1+\sqrt{\alpha})^\mu - (1-\sqrt{\alpha})^\mu}\right)^2\right)^{-1}.$$

The constant γ *is the constant of the strengthened Cauchy-inequality (5.4.3), the parameter* b *the constant of the eigenvalue estimate (5.4.1). The parameter* α *is the smallest positive solution of the polynomial equation*

$$1-\gamma^2 = tb + \left(\frac{(1+\sqrt{t})^\mu + (1-\sqrt{t})^\mu}{2\sum_{s=1}^{\mu}(1+\sqrt{t})^{\mu-s}(1-\sqrt{t})^{s-1}}\right)^2. \tag{5.4.9}$$

Proof: The proof can be found in [6]. □
We describe now the algorithm in order to solve a linear system with the matrix $\tilde{C}_{k-1,r,\mu}^c$ (5.4.6). From (5.4.4), we can deduce

$$P_\mu(t) = \sum_{j=0}^{\mu} a_j t^j,$$

where $a_0 = 1$ ($P_\mu(0) = 1$). Hence, we obtain

$$\begin{aligned}
(\tilde{C}_{k-1,r,\mu}^c)^{-1} &= (I - P_\mu(\tilde{C}_{k-1,r,\mu}^{-1}K_{k-1}))K_{k-1}^{-1} \\
&= \left(I - \sum_{j=0}^{\mu} a_j(\tilde{C}_{k-1,r,\mu}^{-1}K_{k-1})^j\right)K_{k-1}^{-1} \\
&= -\sum_{j=1}^{\mu} a_j(\tilde{C}_{k-1,r,\mu}^{-1}K_{k-1})^j K_{k-1}^{-1} \\
&= -\tilde{C}_{k-1,r,\mu}^{-1}(a_1 + K_{k-1}\tilde{C}_{k-1,r,\mu}^{-1}(a_2 + \dots \\
&\quad \dots + K_{k-1}\tilde{C}_{k-1,r,\mu}^{-1}(a_{\mu-1} + a_\mu K_{k-1}\tilde{C}_{k-1,r,\mu}^{-1})\dots)).
\end{aligned}$$

Thus, a linear system with the matrix $\tilde{C}_{k-1,r,\mu}^c$ can be solved by μ linear systems solves with the matrix $\tilde{C}_{k-1,r,\mu}$.

5.4.2 Application to $-x^2u_{yy} - y^2u_{xx} = g$.

We apply now this theory to problem (4.2.5). By Theorem 5.26, the constant in the strengthened Cauchy-inequality (5.4.3) can be estimated by

$$\gamma^2 \leq \frac{95}{176}.$$

Thus, we have

$$\frac{1}{\sqrt{1-\gamma^2}} = \frac{4\sqrt{11}}{9} < 2.$$

Using (5.4.8), $\mu = 2$ can be chosen. Hence, by

$$T_\mu(x) = T_2(x) = 2x^2 - 1,$$

the polynomial

$$P_{2,\frac{2}{1+\alpha}}(t) = \left(1 - \frac{2t}{1+\alpha}\right)^2 \tag{5.4.10}$$

is obtained. Furthermore, we have to ensure relation (5.4.1). Using Theorem 5.33, we have the following two eigenvalue estimates between the matrices $\mathfrak{C}_{\mathbb{W}_k}$ (5.3.48) and $K_{\mathbb{W}_k}$

$$\begin{aligned} \lambda_{min}\left({\mathfrak{C}_{\mathbb{W}_k}}^{-1} K_{\mathbb{W}_k}\right) &\geq c_{18}, \\ \lambda_{max}\left({\mathfrak{C}_{\mathbb{W}_k}}^{-1} K_{\mathbb{W}_k}\right) &\leq c_{19} \end{aligned}$$

for all $k \in \mathbb{N}$, where $c_{18} = 1 - \frac{2}{11}\sqrt{11}$ and $c_{19} = 1 + \frac{2}{11}\sqrt{11}$. Equivalent to this fact is

$$\begin{aligned} \lambda_{min}\left({K_{\mathbb{W}_k}}^{-1} \mathfrak{C}_{\mathbb{W}_k}\right) &\geq c_{19}^{-1}, \\ \lambda_{max}\left({K_{\mathbb{W}_k}}^{-1} \mathfrak{C}_{\mathbb{W}_k}\right) &\leq c_{18}^{-1}. \end{aligned}$$

We introduce the matrix

$$\tilde{C}_{22,l} = c_{19}\mathfrak{C}_{\mathbb{W}_l} \tag{5.4.11}$$

for $l = 2, \ldots, k$. Hence, the relation

$$(K_{22,l}\underline{v}, \underline{v}) = (K_{\mathbb{W}_l}\underline{v}, \underline{v}) \leq (\tilde{C}_{22,l}\underline{v}, \underline{v}) \leq \frac{c_{19}}{c_{18}}(K_{22,l}\underline{v}, \underline{v})$$

is valid for all $\underline{v} \in \mathbb{R}^m$, e.g. (5.4.1) is satisfied with

$$\tilde{b} = -1 + \frac{c_{19}}{c_{18}} = \frac{4}{7}\sqrt{11} + \frac{8}{7} < \frac{9153}{2992}. \tag{5.4.12}$$

With $b = \frac{9153}{2992}$ and $\gamma^2 = \frac{95}{176}$, the smallest positive solution of (5.4.9) is

$$\alpha = \frac{1}{17}.$$

Thus, we choose

$$P_{2,\frac{17}{9}}(t) = \left(1 - \frac{17}{9}t\right)^2. \tag{5.4.13}$$

We summarize these observations in the next theorem.

THEOREM 5.39. *Let $\tilde{C}_{k,r,\mu}$ be the matrix of Definition 5.37, where $\tilde{C}_{22,l}$, $l = 2, \ldots, k$, is defined in (5.4.11) and the polynomial $P_{2,\frac{17}{9}}(t)$ is defined via relation (5.4.13). Then,*

$$\begin{aligned} \lambda_{min}\left(\tilde{C}_{k,r,\mu}^{-1} K_k\right) &\geq c_{20}, \\ \lambda_{max}\left(\tilde{C}_{k,r,\mu}^{-1} K_k\right) &\leq 1 \end{aligned}$$

hold for all $k \in \mathbb{N}$, where

$$c_{20} = (1 - \gamma^2)\frac{4\alpha^2}{\alpha^2(4b+1) + 1 + 2\alpha} = \frac{17}{3105} \approx 0.00547.$$

5.5 Other multiplicative multi-level algorithms

In the previous sections, the discretization of the degenerated problem (4.2.1) via finite elements is considered. Now, we will focus additionally on the finite difference discretizations of (4.2.2) and (4.2.1), too. We will derive algorithms which are more efficient in numerical experiments as the algorithms described in sections 5.3 and 5.4. However, we cannot prove a mesh-size independent convergence result.

5.5.1 Multi-grid for finite element discretizations

The theory of Theorems 5.39 and 5.36, i.e. the condition number of the AMLI preconditioner (5.4.5) and the convergence rate of the multi-grid Algorithm 5.3 $MULT$, is confirmed in numerical experiments, cf. section 5.8. However, the absolute number of iterations can be reduced. Furthermore, if the number μ of cycles per level for the algorithm $MULT$, or the degree μ of the polynomial iteration for the AMLI preconditioner is equal to one, the numerical results are not satisfactory. The usual multi-grid algorithm, cf. Remark 5.4, is very similar to Algorithm 5.3. Important for such an algorithm is the choice of a proper smoother S which operates on the space $\mathbb{V}_l$, $l = 2, \ldots, k$. A simple Jacobi or Gauß-Seidel smoother cannot handle the anisotropies of the differential operator in (5.1.1). It is referred to the preprint [11] for numerical examples. Therefore, more appropriated smoothers have to be considered.
The first one is the product of the line Gauß-Seidel smoother in x-direction $S_{x,k}$, see [11], [40], whose error transion operator is given by

$$S_{x,k} = I - 2n^2 \left(D_4 \otimes \bar{T}_2 + T_2 \otimes D_4 \right)^{-1} K_k$$

and the line Gauß-Seidel smoother $S_{y,k}$ in y-direction, whose error transion operator was given by

$$S_{y,k} = I - 2n^2 \left(D_4 \otimes T_2 + \bar{T}_2 \otimes D_4 \right)^{-1} K_k.$$

The matrix $\bar{T}_2$ denotes the lower triangular part of T_2 (3.4.7), i.e.

$$\bar{T}_2 = \frac{1}{2} \begin{bmatrix} 2 & 0 & \ldots & 0 \\ -1 & 2 & 0 & \ddots \\ \ddots & \ddots & \ddots & 0 \\ & 0 & -1 & 2 \end{bmatrix}.$$

Then, let

$$S_{3,k} = S_{x,k} S_{y,k} \tag{5.5.1}$$

which is used as pre-smoother. As post-smoother, we use $S_{3,k}^T$.
The second considered smoother is an extension of the smoother $S_{0,k} \leftrightarrow \underline{S}_{0,k}$ (5.3.49) operating on $\mathbb{W}_k$ to the space $\mathbb{V}_k$. More precisely, a matrix L is defined by setting all that off-diagonal entries of the matrix K_k, cf. (4.2.6), to 0 which are relatively small in comparison to the main diagonal entries of that row and column. Let

$$K_k = \left[a_{ij}^k \right]_{i,j=1}^m, \quad m = (n-1)^2.$$

Then, the matrix

$$L_k = \left[l_{ij}^k\right]_{i,j=1}^m \tag{5.5.2}$$

is defined with the entries

$$l_{ij}^k = \begin{cases} a_{ij}^k & \text{if} \quad 4 \mid a_{ij}^k \mid \geq \max\{a_{ii}^k, a_{jj}^k\} \\ 0 & \text{else} \end{cases} .$$

We introduce the smoother $S_{1,k}$ by its error transion operator, i.e.

$$S_{1,k} = I - \omega L_k^{-1} K_k. \tag{5.5.3}$$

This construction is very similar to that of $S_{0,k}$ (5.3.49), compare the definition of the matrices $\mathfrak{C}_{s,ij}$, $s = 1, 2$, in (5.3.35), (5.3.39) and (5.3.46).
The third smoother is the ILU-smoother. Its error transion operator is defined as

$$S_{2,k} = I - \omega (D_k + U_k)^{-1} D_k (D_k + U_k^T)^{-1} K_k, \tag{5.5.4}$$

where D_k is a diagonal matrix, $U_k = [u_{rs}^k]_{r,s=1}^m$ is a strongly upper triangular matrix and $K_k = [a_{rs}^k]_{r,s=1}^m$. The matrix $(D_k + U_k^T) D_k^{-1} (D_k + U_k)$ is called the incomplete LU-decomposition (ILU) of the (symmetric) matrix K_k, if the following conditions are fulfilled:

- If $a_{rs}^k = 0$, then $u_{rs}^k = 0$.
- Let $K_k = (D_k + U_k^T) D_k^{-1} (D_k + U_k) + B_k$, where $B_k = [b_{rs}^k]_{r,s=1}^m$. If $a_{rs}^k \neq 0$, then $b_{rs}^k = 0$.

Using these conditions, the ILU-decomposition can be computed for matrices with 5-point stencil structure as K_k (4.2.6), [42]. One obtains

$$u_{rs}^k = \begin{cases} a_{rs}^k & \text{if} \quad 1 \leq r < s \leq m, \\ 0 & \text{if} \quad m \geq r \geq s \geq 1 \end{cases}$$

for the entries of U_k. Moreover, the entries of the matrix $D_k = \operatorname{diag}[\mathfrak{m}]$, $\mathfrak{m} = [\mathfrak{m}_{(i,j)}]_{(i,j)=(1,1)}^{(n-1,n-1)}$ can be computed recursively by the relations

$$\begin{aligned} \mathfrak{m}_{(1,1)} &= a_{11}^k, \\ \mathfrak{m}_{(i,1)} &= a_{ss}^k, \quad s = (n-1)(i-1) + 1, \quad i \geq 2, \\ \mathfrak{m}_{(1,j)} &= a_{jj}^k, \quad j \geq 2, \\ \mathfrak{m}_{(i,j)} &= a_{ss}^k - \frac{(j^2 + 1/6)^2}{\mathfrak{m}_{(i-1,j)}} - \frac{(i^2 + 1/6)^2}{\mathfrak{m}_{(i,j-1)}}, \quad s = (n-1)(i-1) + j, \quad i, j \geq 2. \end{aligned}$$

5.5.2 Multi-grid preconditioner

If a mesh-size independent convergence rate can be proved in the energy norm, a multi-grid preconditioner can be built. Then, the condition number of the preconditioned system is bounded by a constant independent of the level number k. We write the Algorithm 5.3 in order to solve

$$K_k \underline{u}_k = \underline{g}_k$$

in terms of matrices. Let $\underline{u}_{0,k}$ be the initial value. The new iterate

$$\underline{u}_{1,k} = M_{k,S,\mu}\underline{u}_{0,k} + (I - M_{k,S,\mu})K_k^{-1}\underline{g}_k \tag{5.5.5}$$

will be computed as follows:

- Pre-smoothing: Do $\underline{u}_{0,1,k} = S_{k,pre}^{\nu}\underline{u}_{0,k} + (I - S_{k,pre}^{\nu})K_k^{-1}\underline{g}_k$ with

$$S_{k,pre} = I - \omega\tilde{K}_{k,pre}^{-1}K_k. \tag{5.5.6}$$

- Calculation and restriction of the defect: Set

$$\underline{d}_{k-1} = Q_k^{k-1}(\underline{g}_k - K_k\underline{u}_{0,1,k})$$

  with the finite element restriction matrix Q_k^{k-1}.

- Solve the coarse grid system $K_{k-1}\underline{w}_{k-1} = \underline{d}_{k-1}$ by a direct solver for $k = 2$ and by

$$\underline{u}_{j,k-1} = M_{k-1,S,\mu}\underline{u}_{j-1,k-1} + (I - M_{k-1,S,\mu})K_{k-1}^{-1}\underline{d}_{k-1}, \quad j = 1, \ldots, \mu,$$

  for $k > 2$. Set $\underline{w}_{k-1} = \underline{u}_{\mu,k-1}$ (the initial vector is the vector $[0, \ldots, 0]^T$).

- Interpolation and correction of the defect: Set $\underline{u}_{0,2,k} = \underline{u}_{0,1,k} + Q_{k-1}^k\underline{w}_k$, where $Q_{k-1}^k = (Q_k^{k-1})^T$.

- Post-smoothing: Do $\underline{u}_{1,k} = S_{k,pre}^{\nu}\underline{u}_{0,k} + (I - S_{k,pre}^{\nu})K_k^{-1}\underline{g}_k$ with

$$S_{k,post} = I - \omega\tilde{K}_{k,post}^{-1}K_k. \tag{5.5.7}$$

Thus, one iteration of the multi-grid algorithm can be interpreted as one iteration of a simple iterative method, i.e.

$$\underline{u}_{1,k} = \underline{u}_{0,k} - C_{k,S,\mu}^{-1}(K_k\underline{u}_{0,k} - \underline{g}_k)$$

with the preconditioner $C_{k,S,\mu}^{-1} = (I - M_{k,S,\mu})K_k^{-1}$. However, more efficient iterative methods with preconditioning are introduced in subsection 2.1. One example is the preconditioned conjugate gradient method. The following result can be proved.

THEOREM 5.40. *Let us assume that the following assumptions are satisfied:*

(i) Let K_l, $l = 1, \ldots, k$, be symmetric and positive definite matrices.

(ii) For $l = 2, \ldots, k$, the matrices $S_{l,pre}$ (5.5.6) and $S_{l,post}$ (5.5.7) are adjoint in the K_l scalar product:

$$(S_{l,pre}\underline{u}, \underline{v})_{K_l} = (\underline{u}, S_{l,post}\underline{v})_{K_l} \quad \forall \underline{u}, \underline{v}.$$

(iii) *For $l = 2, \ldots, k$, the restriction and interpolation operators Q_l^{l-1} and Q_l^{l-1} are adjoint in the Euclidian scalar product*

$$\left(Q_l^{l-1}\underline{u}, \underline{v}\right) = \left(\underline{u}, Q_{l-1}^{l}\underline{v}\right) \quad \forall \underline{u}, \underline{v}.$$

(iv) *The mg-operator in (5.5.5) satisfies the estimate*

$$\sup_{\underline{u}_{0,k} \neq \mathbf{0}} \frac{\left(\underline{u}_{1,k}, \underline{u}_{1,k}\right)_{K_k}}{\left(\underline{u}_{0,k}, \underline{u}_{0,k}\right)_{K_k}} \leq \sigma^2$$

for all $k \in \mathbb{N}$ with $0 < \sigma < 1$.

Furthermore, we define the preconditioner

$$\bar{C}_{k,S,\mu,j}^{-1} = (I - M_{k,S,\mu}^{j})K_k^{-1}, \quad j \in \mathbb{N}. \tag{5.5.8}$$

This preconditioner is symmetric and positive definite. Moreover, there

$$\begin{aligned} \lambda_{min}\left(\bar{C}_{k,S,\mu,j}^{-1}K_k\right) &= 1 - \sigma^j, \\ \lambda_{max}\left(\bar{C}_{k,S,\mu,j}^{-1}K_k\right) &= \begin{cases} 1 & j \quad \textit{even} \\ 1 + \sigma^j & j \quad \textit{odd} \end{cases} \end{aligned}$$

hold.

Proof: The proof is a special case of Theorem 2.1. in [51] with $C_k = K_k$, see also [38], Theorems 6.5. and 6.6. □

We note that the finite element restriction and interpolation operators Q_l^{l-1} and Q_{l-1}^{l} fulfill assumption (iii). By Remark 5.2, the matrices K_l are symmetric and positive definite. Thus, assumption (i) holds. For the smoothers $S_{k,post/pre} = I - \omega\tilde{K}_{k,post/pre}^{-1}K_k$, see (5.5.6), (5.5.7), of the Richardson type with

$$\tilde{K}_{k,post} = \tilde{K}_{k,pre}^{T}, \tag{5.5.9}$$

one obtains

$$\begin{aligned} \left(S_{l,pre}\underline{u}, \underline{v}\right)_{K_l} &= \left((I - \omega\tilde{K}_{l,pre}^{-1}K_l)\underline{u}, \underline{v}\right)_{K_l} \\ &= \left((K_l - \omega K_l\tilde{K}_{l,pre}^{-1}K_l)\underline{u}, \underline{v}\right) \\ &= \left(K_l\underline{u}, (I - \omega\tilde{K}_{l,pre}^{-T}K_l)\underline{v}\right) \\ &= \left(\underline{u}, S_{l,post}\underline{v}\right)_{K_l} \end{aligned}$$

which means that assumption (ii) is satisfied. By the symmetry of the matrix $\mathfrak{C}_{\mathbb{W}_l}$ (5.3.48), the smoother $S_{0,l}$ (5.3.49) fulfills relation (5.5.9) and so assumption (ii). By same arguments, assumption (ii) is valid for the smoothers $S_{1,l}$ (5.5.3) and $S_{2,l}$ (5.5.4). However, the smoother $S_{3,l}$ (5.5.1) does not fulfill relation (5.5.9). In this case, set $S_{l,pre} = S_{3,l}$ and $S_{l,post} = S_{3,l}^{T}$. Then, $S_{l,pre}$ is the product of forwards line Gauß-Seidel smoother in x- and y-direction, whereas $S_{l,post}$ is the product backwards line Gauß-Seidel smoother in y- and x-direction. Moreover by Theorem 5.36, assumption (iv) of Theorem 5.40 is valid for the smoother $S_{0,l} \hookrightarrow \underline{S}_{0,l}$ with $\nu \geq 3$ and $\mu = 3$. Therefore, the following theorem has been proved.

THEOREM 5.41. *The symmetric and positive definite matrix $\bar{C}_{k,S_{0,k},3,j}$, see (5.5.8), satisfies $\kappa\left(\bar{C}^{-1}_{k,S_{0,k},3,j}K_k\right) \leq c$ with a constant c independent of the level number k for all $j \in \mathbb{N}$.*

In the following, we consider the case $j = 1$ in (5.5.8), i.e. one iteration multi-grid per preconditioning step, only. Then, the last index in $\bar{C}_{k,S,\mu,j}$ is omitted, e.g. $\bar{C}_{k,S,\mu} := \bar{C}_{k,S,\mu,1}$.

5.5.3 Multi-grid for finite difference discretizations

In this subsection, the finite difference discretizations of problems (4.2.1) and (4.2.2) are investigated. As result of this discretization, the systems

$$C_3\underline{u}_k = \underline{g}_k \quad \text{and} \tag{5.5.10}$$
$$C_1\underline{u}_k = \underline{g}_k \tag{5.5.11}$$

have to be solved, cf. (3.4.17) for the definition of C_3 and (3.4.3) for the definition of C_1. Similar as $K_k\underline{u}_k = \frac{1}{2n^2}C_4\underline{u}_k = \underline{g}_k$, the systems (5.5.10) and (5.5.11) will be solved by a standard multi-grid algorithm.
Via the matrices K_k (4.2.6), or C_4, cf. relations (4.2.9) and (3.4.18), preconditioners for $A_{\mathcal{R}_2} = \text{blockdiag}\,[\mathfrak{A}_i]_{i=1}^4$ (3.3.4) can be built, where the condition numbers of the preconditioned systems grow as $(1 + \log p)$. The reason of this logarithmic term is the condition number estimate $\kappa\left({C_4}^{-1}\mathfrak{A}_1\right) \preceq (1 + \log p)$, cf. Theorem 3.11. Since $\kappa\left({C_1}^{-1}\mathfrak{A}_1\right) = \mathcal{O}(1)$, cf. Lemma 3.5, a spectrally equivalent preconditioner for $A_{\mathcal{R}_2}$ (3.3.4) can be developed via the matrix

$$C_1 = D_3 \otimes T_1 + T_1 \otimes D_3.$$

Therefore, it is important to derive a fast solver for C_1, the discretization of problem (4.2.2) by finite differences. Let $S_{2,k}$ be the ILU-smoother for C_1 and let $S_{3,k}$ be the product of the line Gauß-Seidel smoother in x-direction and the line Gauß-Seidel smoother in y-direction for C_1. For reasons of a simple notation, we denote these smoothers for C_1 with $S_{2,k}$ and $S_{3,k}$ as well as the smoothers $S_{2,k}$ (5.5.4) and $S_{3,k}$ (5.5.1) for $C_4 \leftrightarrow K_k$. The first index of S indicates only the construction method for the smoother, i.e. 2 for the ILU smoother and 3 for the product of the line Gauß-Seidel smoothers. The system (5.5.11) is solved by a standard multi-grid algorithm for finite difference discretizations with bilinear interpolation. The used smoothers are $S_{2,k}$ (as pre- and post-smoother), or $S_{3,k}$ as pre-smoother and $S^T_{3,k}$ as post-smoother. The corresponding multi-grid operator is denoted by $\breve{C}_{k,S,\mu}$, where S denotes the kind of smoother, the integer μ the number of cycles per level and k denotes the level number. Moreover, we define

$$\breve{C}_{k,S,\mu} = (I - \breve{M}_{k,S,\mu})C_1^{-1} \tag{5.5.12}$$

as the corresponding multi-grid preconditioner for C_1 with one iteration multi-grid.

5.6 BPX preconditioner

5.6.1 Definition of the preconditioners

Recall the finite element discretization of problem (4.2.3) in subsection 4.2.2:
Find $u \in \mathbb{V}_k$ such that

$$\int_\Omega \left(\omega^2(y)u_x v_x + \omega^2(x)u_y v_y\right)\ \mathrm{d}(x,y) = \int_\Omega fv\ \mathrm{d}(x,y) \tag{5.6.1}$$

holds for all $v \in \mathbb{V}_k$ with a weight function $\omega(\xi)$ satisfying Assumption 5.27.
For the efficient solution of systems of linear equations arising from discretizations of uniformly elliptic problems by finite elements, Bramble, Pasciak, and Xu have developed a preconditioner, [21], which was called the BPX preconditioner. For this preconditioner, the spectral equivalence to the original stiffness matrix can be shown. Later, this preconditioner has been improved by the multiple diagonal scaling version, [81]. As mentioned in section 5.2, cf. [10], a BPX preconditioner with multiple diagonal scaling does not show good numerical results in order to solve $K_k\underline{u} = \underline{g}_k$, the system of linear algebraic equations resulting from the finite element discretization of (5.6.1). One reason is that this preconditioner cannot handle the anisotropies resulting from the degenerated elliptic operator. However, with a modification, the so called multiple tridiagonal scaling BPX (MTS-BPX), this behaviour of the BPX preconditioner can be improved, [12]. In subsection 5.5.1, several smoothers of Richardson-type are considered.
One smoother is, cf. (5.5.3),

$$S_{1,k} = I - \omega L_k^{-1} K_k.$$

In this smoother, the matrix L_k is a preconditioner for K_k which can handle anisotropies. The idea is to apply the matrix L_k as "scaling" on each level instead of a diagonal scaling. We expect a stabilization of the BPX preconditioner. The following MTS-BPX preconditioner is now defined. Let Q_l^k, $l = 1, \ldots, k$ be the basis transformation matrix from the basis $\{\phi_{ij}^l\}_{i,j}^{n_l-1} \in \mathbb{V}_l$ to the basis $\{\phi_{ij}^k\}_{i,j=1}^{n_k-1} \in \mathbb{V}_k$, where $n_l = 2^l$. Let Q_k^l be the transposed operator. Furthermore, let L_k be the matrix (5.5.2). Then, we define the preconditioner

$$\hat{C}_k^{-1} = \sum_{l=1}^{k} Q_l^k L_l^{-1} Q_k^l. \tag{5.6.2}$$

This preconditioner is called the MTS-BPX preconditioner for K_k.
Choosing the ILU-decomposition of the matrix K_k, another additive multi-level preconditioner can be defined. Let, cf. (5.5.4),

$$\mathcal{L}_k^{-1} = (D_k + U_k)^{-1} D_k (D_k + U_k^T)^{-1}$$

be the inverse ILU-decomposition of the matrix K_k. Then, we define the ILU-BPX preconditioner

$$\hat{\mathcal{C}}_k^{-1} = \sum_{l=1}^{k} Q_l^k \mathcal{L}_l^{-1} Q_k^l. \tag{5.6.3}$$

As for the preconditioner $\hat{C}_k$ (5.6.2), we expect a better handling of the anisotropies.
For the correct analytical definition of the MTS-BPX preconditioner $\hat{C}_k$, we recall the notation of subsection 4.2.2 and introduce some new notation. Let $\mathbb{V}_k = \operatorname{span}\left\{\phi_{ij}^k\right\}_{i,j=1}^{n_k-1}$, where k denotes the level number and $n_k = 2^k$. Moreover, let $k' \leq k$. The domain Ω is decomposed into overlapping stripes $\hat{\Omega}_j^k$, i.e.

$$\overline{\Omega} = \bigcup_{j=1}^{n_k-1} \hat{\Omega}_j^k,$$

where $\hat{\Omega}_j^k = \hat{\Omega}_{j,x}^k \cup \hat{\Omega}_{j,y}^k$ with

$$\begin{aligned}
\hat{\Omega}_{j,x}^k &= \left\{(x,y) \in \mathbb{R}^2, 0 \leq y \leq x, \frac{j-1}{n_k} \leq x \leq \frac{j+1}{n_k}\right\}, \\
\hat{\Omega}_{j,y}^k &= \left\{(x,y) \in \mathbb{R}^2, 0 \leq x \leq y, \frac{j-1}{n_k} \leq y \leq \frac{j+1}{n_k}\right\},
\end{aligned}$$

see Figure 5.2. According to this decomposition, let

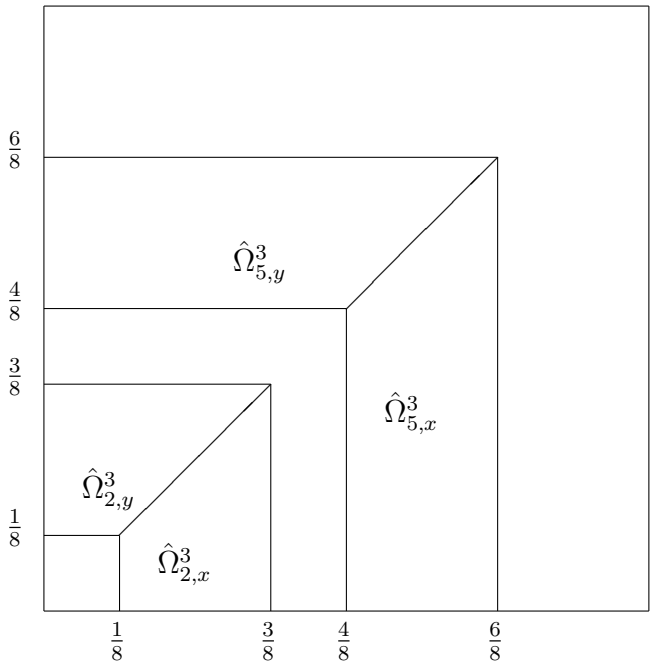

Figure 5.2: Stripes $\hat{\Omega}_j^k$ for $k = 3$ and $j = 2, 5$.

$$\mathbb{V}_j^k = \operatorname{span}\left\{\phi_{ij}^k\right\}_{i=1}^{j-1} \oplus \operatorname{span}\left\{\phi_{ji}^k\right\}_{i=1}^{j} \tag{5.6.4}$$

be the corresponding finite element subspaces to the sub-domains $\hat{\Omega}_j^k$. Note that all shape functions $\phi^k \in \mathbb{V}_j^k$ vanish on the boundary of $\hat{\Omega}_j^k$. The additive Schwarz splitting of the finite element

space $\mathbb{V}_k$, i.e.

$$\mathbb{V}_k = \sum_{k'=1}^{k} \sum_{j=1}^{n_{k'}-1} \mathbb{V}_j^{k'}$$

is considered. Following Zhang, [81], let $\underline{K}_k : \mathbb{V}_k \mapsto \mathbb{V}_k$ and $\underline{K}_{i,k} : \mathbb{V}_i^k \mapsto \mathbb{V}_i^k$ be the operators

$$\begin{aligned} \langle \underline{K}_k u, v \rangle &= a(u,v) \quad \forall u, v \in \mathbb{V}_k, \\ \langle \underline{K}_{i,k} u, v \rangle &= a(u,v) \quad \forall u, v \in \mathbb{V}_i^k. \end{aligned}$$

Moreover, let $P_{i,k'} : \mathbb{V}_k \mapsto \mathbb{V}_i^{k'}$ be the energetic projection and $Q_{i,k'} : \mathbb{V}_k \mapsto \mathbb{V}_i^{k'}$ be the L^2-projection, i.e.

$$\begin{aligned} a(P_{i,k'} u, v) &= a(u,v) \quad \forall v \in \mathbb{V}_i^{k'}, \\ \langle Q_{i,k'} u, v \rangle &= \langle u, v \rangle \quad \forall v \in \mathbb{V}_i^{k'}, \end{aligned}$$

where $u \in \mathbb{V}_k$. Then, the preconditioner $\underline{\hat{C}}_k$ and the k-th level additive Schwarz operator P_k are defined by

$$\underline{\hat{C}}_k^{-1} = \sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}-1} \underline{K}_{i,k'}^{-1} Q_{i,k'}, \tag{5.6.5}$$

$$P_k = \underline{\hat{C}}_k^{-1} \underline{K}_k = \sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}-1} P_{i,k'}. \tag{5.6.6}$$

Note that the matrices K_k (4.2.6) and $\hat{C}_k$ (5.6.2) denote the matrix representations of $\underline{K}_k$ and $\underline{\hat{C}}_k$ by the usual fem-isomorphism. For technical reasons, we investigate the additive Schwarz splitting

$$\mathbb{V}_k = \sum_{k'=1}^{k} \mathbb{U}_1^{k'} \oplus \mathbb{U}_2^{k'}, \tag{5.6.7}$$

where

$$\begin{aligned} \mathbb{U}_1^{k'} &= \mathbb{V}_1^{k'} \oplus \mathbb{V}_3^{k'} \oplus \cdots \oplus \mathbb{V}_{n_{k'}-1}^{k'} \quad \text{and} \\ \mathbb{U}_2^{k'} &= \mathbb{V}_2^{k'} \oplus \mathbb{V}_4^{k'} \oplus \cdots \oplus \mathbb{V}_{n_{k'}-2}^{k'} \end{aligned} \tag{5.6.8}$$

as well (cf. (5.6.4)). Let $\underline{\tilde{K}}_{s,k}$: $\mathbb{U}_s^k \mapsto \mathbb{U}_s^k$, $\tilde{P}_{s,k'}$: $\mathbb{V}_k \mapsto \mathbb{U}_s^{k'}$ and $\tilde{Q}_{s,k'}$: $\mathbb{V}_k \mapsto \mathbb{U}_s^{k'}$ be the operators

$$\begin{aligned} \langle \underline{\tilde{K}}_{s,k} u, v \rangle &= a(u,v) \quad \forall u, v \in \mathbb{U}_s^k, \\ a\left(\tilde{P}_{s,k'} u, v\right) &= a(u,v) \quad \forall u \in \mathbb{V}_k, v \in \mathbb{U}_s^{k'}, \\ \langle \tilde{Q}_{s,k'} u, v \rangle &= \langle u, v \rangle \quad \forall u \in \mathbb{V}_k, v \in \mathbb{U}_s^{k'}, \end{aligned}$$

where $s = 1, 2$. Thus, the preconditioner $\underline{\hat{C}}_k$ (5.6.5) and the k-th level additive Schwarz operator $\hat{P}_k$ can be obtained as multi-level additive Schwarz preconditioner and projection operator corresponding to (5.6.7).

LEMMA 5.42. *The relations*

$$\underline{\hat{C}}_k^{-1} = \sum_{k'=1}^{k} \sum_{s=1}^{2} \underline{\tilde{K}}_{s,k'}^{-1} \tilde{Q}_{s,k'} \quad \textit{and} \tag{5.6.9}$$

$$P_k = \sum_{k'=1}^{k} \sum_{s=1}^{2} \tilde{P}_{s,k'} \tag{5.6.10}$$

are valid.

Proof: Note that $a(u,v) = 0$ and $\langle u, v \rangle = 0$ for all $u \in \mathbb{V}_i^{k'}$ and $v \in \mathbb{V}_j^{k'}$ with $|i-j| \geq 2$. Thus, the sums in (5.6.8) are orthogonal sums with respect to $a(\cdot,\cdot)$ and $\langle \cdot, \cdot \rangle$. Hence, the L^2 and the energetic projection from $\mathbb{V}_k$ onto $\mathbb{U}_{s,k'}$ is the sum of the projections onto $\mathbb{V}_{2i-2+s}^{k'}$, $i = 1, \ldots, \frac{n_{k'}}{2} + 1 - s$, i.e.

$$\tilde{Q}_{s,k'} u = \sum_{i=1}^{\frac{n_{k'}}{2}+1-s} Q_{2i-2+s} u, \tag{5.6.11}$$

$$\tilde{P}_{s,k'} u = \sum_{i=1}^{\frac{n_{k'}}{2}+1-s} P_{2i-2+s} u$$

hold for all $u \in \mathbb{V}_{k'} \subset \mathbb{V}_k$. Therefore, relation (5.6.10) has been proved. Moreover, let

$$u = \sum_{i=1}^{\frac{n_{k'}}{2}+1-s} u_{2i-2+s}, \quad u_j \in \mathbb{V}_j^{k'}, u \in \mathbb{U}_s^{k'}, s = 1, 2.$$

Since $a(u_i, u_j) = 0$ for all $u_j \in \mathbb{V}_j^{k'}$ and $u_i \in \mathbb{V}_i^{k'}$ with $|i-j| \geq 2$,

$$\underline{\tilde{K}}_{s,k'} u = \sum_{i=1}^{\frac{n_{k'}}{2}+1-s} \underline{\tilde{K}}_{2i-2+s,k'} u_{2i-2+s} \quad \text{or} \quad \left(\underline{\tilde{K}}_{s,k'}\right)^{-1} u = \sum_{i=1}^{\frac{n_{k'}}{2}+1-s} \left(\underline{\tilde{K}}_{2i-2+s,k'}\right)^{-1} u_{2i-2+s}$$

follows. Together with (5.6.11) and (5.6.8), the assertion (5.6.9) has been proved. □

5.6.2 Proof of the upper eigenvalue estimate

We prove now the estimate $\lambda_{max}(P_k) \leq c\,k$ with a constant c independent of the mesh-size h. Two proofs are given.

The first proof is similar to the proof of Zhang for the upper eigenvalue bound of the MDS-BPX preconditioned system matrix given in [81]. Zhang has proved that the condition number of the preconditioned system is bounded by a constant independent of the level number, if the bilinear form $a(\cdot,\cdot)$ is uniformly elliptic and bounded. Using the techniques of Zhang, we can only prove the result $\lambda_{max}\left(\hat{C}_k^{-1} K_k\right) = \lambda_{max}(P_k) \leq c\,k$ for the MTS-BPX preconditioner. The second

proof uses the multi-level additive Schwarz splitting $\mathbb{V}_k = \sum_{k'=1}^{k} \mathbb{U}_1^{k'} \oplus \mathbb{U}_2^{k'}$ (5.6.7). Using this space splitting, the result $\lambda_{max}(P_k) \leq c\,k$ can be established by a short proof. This proof requires the positive definiteness of the bilinear form $a(\cdot,\cdot)$ only. The Zhang-like proof is given in order to show that $\lambda_{max}(P_k) \leq c$ cannot be concluded by a more rigorous estimate. Numerical experiments indicate that the upper eigenvalue of P_k grows as the level number k.
Now, we start with the first proof. For this aim, the following lemma is useful. (Recall Figure 5.1 for the definition of the triangles $\tau_{ij}^{1,k}$ and $\tau_{ij}^{2,k}$.)

LEMMA 5.43. *For weight functions satisfying Assumption 5.27, the estimate*

$$\int_{\tau_{rs}^{1,k}} \omega^2(y)\,\mathrm{d}(x,y) \asymp \int_{\tau_{rs}^{2,k}} \omega^2(y)\,\mathrm{d}(x,y) \tag{5.6.12}$$

is valid for all $r, s \in \mathbb{N}_0$.

Proof: By the monotony of the weight function, one easily checks

$$\int_{\tau_{rs}^{2,k}} \omega^2(y)\,\mathrm{d}(x,y) \leq \int_{\tau_{rs}^{1,k}} \omega^2(y)\,\mathrm{d}(x,y). \tag{5.6.13}$$

By Lemma 5.29 on page 59, we have

$$0 \leq \left(\omega\left(y+\frac{1}{4n_k}\right)\right)^2 \leq c(\omega(y))^2 \quad \forall y \geq \frac{1}{2n_k}.$$

Integration with respect to the variable y gives

$$\begin{aligned}\int_{\frac{4j+2}{4n_k}}^{\frac{4j+3}{4n_k}} \left(\omega\left(y+\frac{1}{4n_k}\right)\right)^2 \mathrm{d}y &\leq c\int_{\frac{4j+2}{4n_k}}^{\frac{4j+3}{4n_k}} \omega^2(y)\,\mathrm{d}y \quad \forall j \in \mathbb{N}_0, \quad \text{or,}\\ \int_{\frac{4j+3}{4n_k}}^{\frac{4j+4}{4n_k}} \omega^2(y)\,\mathrm{d}y &\leq c\int_{\frac{4j+2}{4n_k}}^{\frac{4j+3}{4n_k}} \omega^2(y)\,\mathrm{d}y \quad \forall j \in \mathbb{N}_0.\end{aligned}$$

By integration with respect to the variable x from $\frac{4i}{4n_k}$ to $\frac{4i+1}{4n_k}$, one concludes $(4n_k = 2^{k+2})$

$$\begin{aligned}\int_{\mathcal{E}_{4i,4j+3}^{k+2}} \omega^2(y)\,\mathrm{d}(x,y) &\leq c\int_{\mathcal{E}_{4i,4j+2}^{k+2}} \omega^2(y)\,\mathrm{d}(x,y) \quad \forall i,j \in \mathbb{N}_0\\ &= c\int_{\mathcal{E}_{4i+3,4j+2}^{k+2}} \omega^2(y)\,\mathrm{d}(x,y) \quad \forall i,j \in \mathbb{N}_0.\end{aligned} \tag{5.6.14}$$

For the last estimate, it is used that the integrand does not depend on the variable x. Note that $\mathcal{E}_{4i+3,4j+2}^{k+2} \subset \tau_{ij}^{2,k}$, cf. Figure 5.3. Thus, the inequality

$$\int_{\mathcal{E}_{4i+3,4j+2}^{k+2}} \omega^2(y)\,\mathrm{d}(x,y) \leq \int_{\tau_{ij}^{2,k}} \omega^2(y)\,\mathrm{d}(x,y) \tag{5.6.15}$$

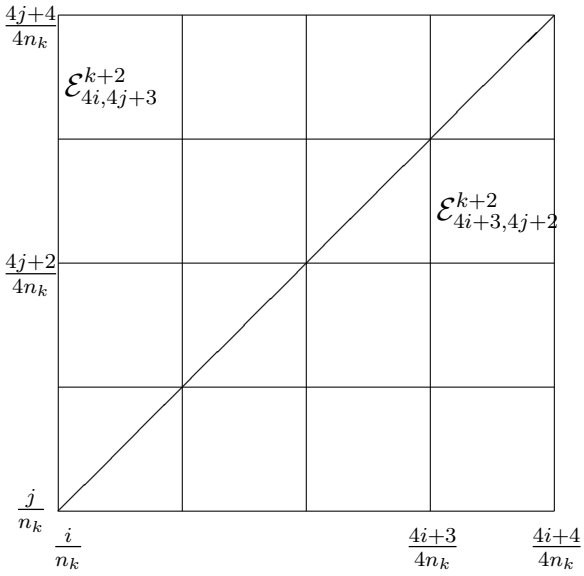

Figure 5.3: Notation for $\mathcal{E}^k_{ij} = \overline{\tau^{1,k}_{ij} \cup \tau^{2,k}_{ij}}$, $n_k = 2^k$.

holds for all $i, j \in \mathbb{N}_0$. Moreover, by $\mathcal{E}^{k+2}_{4i,4j+2} \subset \tau^{1,k}_{ij}$ and the monotony of the weight function, one easily deduces

$$8 \int_{\mathcal{E}^{k+2}_{4i,4j+3}} \omega^2(y)\, \mathrm{d}(x,y) \geq \int_{\tau^{1,k}_{ij}} \omega^2(y)\, \mathrm{d}(x,y). \tag{5.6.16}$$

Combining the estimates (5.6.14), (5.6.15), and (5.6.16), one checks

$$\int_{\tau^{1,k}_{ij}} \omega^2(y)\, \mathrm{d}(x,y) \leq 8c \int_{\tau^{2,k}_{ij}} \omega^2(y)\, \mathrm{d}(x,y). \tag{5.6.17}$$

By (5.6.13) and (5.6.17), the assertion follows immediately. □
Equivalent to the estimate (5.6.12) is that

$$\int_{\tau^{u,k}_{rs}} \omega^2(x)\, \mathrm{d}(x,y) \geq c \int_{\mathcal{E}^k_{rs}} \omega^2(x)\, \mathrm{d}(x,y)$$

is valid for $u = 1, 2$, and $r, s \in \mathbb{N}_0$ with a constant c independent of r, s, and k. The main tool in order to estimate the upper eigenvalue of the BPX preconditioner is Lemma 5.46 which is a strengthened Cauchy-inequality of the type

$$\Big(a(u^k_i, u^k_j)\Big)^2 \leq c 2^{|k'-k|} a(u^{k'}_i, u^{k'}_i) a(u^k_j, u^k_j) \tag{5.6.18}$$

for all $u^k_j \in \mathbb{V}^k_j$ and $u^{k'}_i \in \mathbb{V}^{k'}_i$. Our aim is to prove (5.6.18). We split this proof into several lemmata. The first lemma says that the mean value of the weight function $\omega(x)$ over $\tau^{u,k'}_{rs} \cap \hat{\Omega}^k_{j,x}$ can be bounded by the mean value over $\tau^{u,k'}_{rs}$.

LEMMA 5.44. *For $u = 1, 2$, $r, s \in \mathbb{N}_0$, $k' \leq k$, $j \in \mathbb{N}$, the inequalities*

$$\frac{n_{k'}}{n_k} \int_{\tau_{rs}^{u,k'}} \omega^2(y)\, \mathrm{d}(x,y) \geq c \int_{\tau_{rs}^{u,k'} \cap \hat{\Omega}_{j,x}^k} \omega^2(y)\, \mathrm{d}(x,y) \tag{5.6.19}$$

and

$$\frac{n_{k'}}{n_k} \int_{\tau_{rs}^{u,k'}} \omega^2(x)\, \mathrm{d}(x,y) \geq c \int_{\tau_{rs}^{u,k'} \cap \hat{\Omega}_{j,x}^k} \omega^2(x)\, \mathrm{d}(x,y) \tag{5.6.20}$$

are valid.

Proof: For $\overline{\tau_{rs}^{u,k'}} \cap \hat{\Omega}_{j,x}^k = \emptyset$, the assertion is trivial ($c = 0$). We assume that $\overline{\tau_{rs}^{u,k'}} \cap \hat{\Omega}_{j,x}^k \neq \emptyset$. Then, there

$$\underline{c} \frac{r}{n_{k'}} \leq \frac{j}{n_k} \leq \overline{c} \frac{r+1}{n_{k'}} \tag{5.6.21}$$

holds. Now, with (5.6.12) and Assumption 5.27, we estimate

$$\begin{aligned}
\int_{\tau_{rs}^{u,k'}} \omega^2(x)\, \mathrm{d}(x,y) &\overset{(5.6.12)}{\geq} c \int_{\mathcal{E}_{rs}^{k'}} \omega^2(x)\, \mathrm{d}(x,y) \\
&= c \int_{\frac{r}{n_{k'}}}^{\frac{r+1}{n_{k'}}} \int_{\frac{s}{n_{k'}}}^{\frac{s+1}{n_{k'}}} \omega^2(x)\, \mathrm{d}y\, \mathrm{d}x \\
&\overset{\omega(\xi)=\xi^\alpha}{=} c \frac{1}{n_{k'}} \int_{\frac{r}{n_{k'}}}^{\frac{r+1}{n_{k'}}} x^{2\alpha}\, \mathrm{d}x \\
&\geq \frac{c}{n_{k'}} \frac{(r+1)^{2\alpha}}{{n_{k'}}^{2\alpha+1}} = \frac{1}{{n_{k'}}^2} \left(\frac{r+1}{n_{k'}} \right)^{2\alpha}.
\end{aligned} \tag{5.6.22}$$

Moreover, one concludes

$$\begin{aligned}
\int_{\tau_{rs}^{u,k'} \cap \hat{\Omega}_{j,x}^k} \omega^2(x)\, \mathrm{d}(x,y) &\leq \int_{\mathcal{E}_{rs}^{k'} \cap \hat{\Omega}_{j,x}^k} \omega^2(x)\, \mathrm{d}(x,y) \\
&\leq \int_{\frac{s}{n_{k'}}}^{\frac{s+1}{n_{k'}}} \int_{\frac{j-1}{n_k}}^{\frac{j+1}{n_k}} \omega^2(x)\, \mathrm{d}x\, \mathrm{d}y \\
&= \frac{1}{n_{k'}} \int_{\frac{j-1}{n_k}}^{\frac{j+1}{n_k}} x^{2\alpha}\, \mathrm{d}x \\
&\leq \frac{c}{n_{k'}} \frac{j^{2\alpha}}{n_k^{2\alpha+1}} \leq \frac{c}{n_k n_{k'}} \left(\frac{j}{n_k} \right)^{2\alpha}.
\end{aligned} \tag{5.6.23}$$

Using (5.6.21) and (5.6.23), there

$$\int_{\tau_{rs}^{u,k'} \cap \hat{\Omega}_{j,x}^k} \omega^2(x)\, \mathrm{d}(x,y) \leq \frac{c}{n_k n_{k'}} \left(\frac{r+1}{n_{k'}} \right)^{2\alpha} \tag{5.6.24}$$

holds. Combining (5.6.24) and (5.6.22), the inequality (5.6.20) follows immediately. The estimate (5.6.19) can be proved with similar arguments. □
Let $a_{\hat{\Omega}_j^k}$ be the restriction of the bilinear form a to $\hat{\Omega}_j^k$, i.e.

$$a_{\hat{\Omega}_j^k}(u,v) = \int_{\hat{\Omega}_j^k} \left(\omega^2(y)u_x v_x + \omega^2(x)u_y v_y\right)\, \mathrm{d}(x,y).$$

Using Lemma 5.44, the following result can be shown.

LEMMA 5.45. *Let $u_i^{k'} \in \mathbb{V}_i^{k'}$. Then for $k' \leq k$, the estimate*

$$2^{k'-k} a(u_i^{k'}, u_i^{k'}) \geq c\, a_{\hat{\Omega}_j^k}(u_i^{k'}, u_i^{k'})$$

is valid.

Proof: For each triangle $\tau_{rs}^{u,k'} \subset \hat{\Omega}_{i,x}^{k'}$, $(\nabla u_i^{k'})^T$ is constant on $\tau_{rs}^{u,k'}$. Therefore, using the estimates (5.6.19) and (5.6.20) of Lemma 5.44,

$$\begin{aligned}\int_{\tau_{rs}^{u,k'}} \left(\omega^2(y)(u_i^{k'})_x(u_i^{k'})_x + \omega^2(x)(u_i^{k'})_y(u_i^{k'})_y\right) &= \int_{\tau_{rs}^{u,k'}} \omega^2(y)(u_i^{k'})_x^2 + \int_{\tau_{rs}^{u,k'}} \omega^2(x)(u_i^{k'})_y^2 \\ &\geq \frac{cn_k}{n_{k'}} \int_{\tau_{rs}^{u,k'} \cap \hat{\Omega}_{j,x}^k} (u_i^{k'})_x^2 \omega^2(y) + (u_i^{k'})_y^2 \omega^2(x).\end{aligned}$$

By symmetry of the differential operator (4.2.3), the same result is valid for each triangle $\tau_{rs}^{u,k'} \subset \hat{\Omega}_{i,y}^{k'}$. Summation over all triangles $\tau_{rs}^{u,k'} \subset \hat{\Omega}_i^{k'}$ gives

$$\int_{\hat{\Omega}_i^{k'}} \left((u_i^{k'})_x^2 \omega^2(y) + (u_i^{k'})_y^2 \omega^2(x)\right)\, \mathrm{d}(x,y) \geq c\frac{n_k}{n_{k'}} \int_{\hat{\Omega}_j^k} \left((u_i^{k'})_x^2 \omega^2(y) + (u_i^{k'})_y^2 \omega^2(x)\right)\, \mathrm{d}(x,y),$$

or equivalently,

$$a(u_i^{k'}, u_i^{k'}) \geq c\frac{n_k}{n_{k'}} a_{\hat{\Omega}_j^k}(u_i^{k'}, u_i^{k'}) = c2^{k-k'} a_{\hat{\Omega}_j^k}(u_i^{k'}, u_i^{k'})$$

which proves the lemma. □
The next lemma gives a relation for the cosine of the angle between the spaces $\mathbb{V}_i^{k'}$ and $\mathbb{V}_j^k$ with respect to $a(\cdot,\cdot)$ which in general is defined as

$$\gamma_{\mathbb{U},\mathbb{V}} = \sup_{\substack{u \in \mathbb{U} \\ v \in \mathbb{V} \\ u,v \neq 0}} \frac{a(u,v)}{\sqrt{a(u,u)a(v,v)}}. \tag{5.6.25}$$

LEMMA 5.46. *Let $k' \leq k$ and $i \in \{1,\ldots,n_{k'}-1\}$, $j \in \{1,\ldots,n_k-1\}$. Then,*

$$\gamma^2_{\mathbb{V}_i^{k'},\mathbb{V}_j^k} \leq \max\left\{c2^{-\frac{k-k'}{2}}, 1\right\}.$$

Proof: The proof is similar to the proof of Lemma 3.2. in [81]. Let $u_i^{k'} \in \mathbb{V}_i^{k'}$ and $u_j^k \in \mathbb{V}_j^k$. Then, by the usual Cauchy-inequality on $a_{\hat{\Omega}_j^k}(\cdot,\cdot)$ and Lemma 5.45,

$$\begin{aligned}\left(a(u_i^{k'}, u_j^k)\right)^2 &= \left(a_{\hat{\Omega}_j^k}(u_i^{k'}, u_j^k)\right)^2 \\ &\leq a_{\hat{\Omega}_j^k}(u_i^{k'}, u_i^{k'})\, a(u_j^k, u_j^k) \\ &\leq c2^{k'-k} a(u_i^{k'}, u_i^{k'})\, a(u_j^k, u_j^k)\end{aligned}$$

which shows the assertion.□
Following Zhang, [81], let

$$\Theta = \left[\theta_{ij}^{k',k''}\right]_{(i,k'),(j,k'')},$$

where

$$\theta_{ij}^{k',k''} = \gamma^2_{\mathbb{V}_i^{k'},\mathbb{V}_j^{k''}}, \quad 1 \leq k', k'' \leq k.$$

Our aim is to prove an estimate of the type

$$\| \Theta \|_2 \leq ck.$$

For this purpose, the following propositions and lemmata are helpful.

PROPOSITION 5.47. *Let k', k be fixed with $k' \leq k$. If $\theta_{ij}^{k',k} \neq 0$, then*

$$(i-1)2^{k-k'} \leq j \leq (i+1)2^{k-k'}.$$

Proof: By definition, $\phi \in \mathbb{V}_j^k$ satisfies $\operatorname{supp}\, \phi \subset \hat{\Omega}_j^k$. If $\operatorname{int}(\hat{\Omega}_j^k) \cap \operatorname{int}(\hat{\Omega}_i^{k'}) = \emptyset$, then $\theta_{ij}^{k',k} = 0$. By definition of the stripes $\hat{\Omega}_j^k$, the assertion follows. □
Now, we consider one block of the matrix Θ, i.e.

$$\Theta^{k',k''} = \left[\theta_{ij}^{k',k''}\right]_{i=1,j=1}^{n_{k'},n_{k''}}.$$

Then, the following proposition is valid.

PROPOSITION 5.48. *The Frobenius norm of $\Theta^{k',k''}$ can be estimated by a constant, independent of the mesh-size h, i.e.*

$$\| \Theta^{k',k''} \|_F \leq c \quad \textit{for } 1 \leq k', k'' \leq k.$$

Proof: Without loss of generality, let $k' \leq k''$. By Proposition 5.47, each row of $\Theta^{k',k''}$ has maximal $2^{k''-k'+1}+1$ nonzero matrix entries, and each column maximal 2 nonzero matrix entries. Therefore, the total number of nonzero matrix entries is less than or equal to $2^{k''-k'+2}+2$. By Lemma 5.46, $\theta_{ij}^{k',k''} \leq c2^{\frac{k'-k''}{2}}$ holds. Summing up over all $(\theta_{ij}^{k',k''})^2$ gives

$$\| \Theta^{k',k''} \|_F = \sum_{i,j} (\theta_{ij}^{k',k''})^2 \leq c2^{k'-k''}(2^{k''-k'+2}+2) \leq 6c$$

which proves the lemma.□
The following lemma, [81], gives a relation between the Frobenius norm of the block matrix Θ and the Frobenius norm of $\tilde{\Theta}$, where the entries of the matrix $\tilde{\Theta}$ are the Frobenius norms of the blocks of θ.

LEMMA 5.49. *Let Θ be a $n \times n$ block matrix, i.e. $\Theta = [\Theta_{ij}]_{i,j=1}^{n}$. Moreover, let $\tilde{\Theta} = [\| \Theta_{ij} \|_F]_{i,j=1}^{n}$. Then,*

$$\| \tilde{\Theta} \|_F = \| \Theta \|_F .$$

Proof: Let

$$\Theta = \left[\theta_{l_i,l_j}^{i,j} \right]_{(l_i,i);(l_j,j)} .$$

Then,

$$\| \Theta \|_F^2 = \sum_{i,j} \sum_{l_i,l_j} (\theta_{l_i,l_j}^{i,j})^2 .$$

Moreover, $\| \Theta_{ij} \|_F^2 = \sum_{l_i,l_j} (\theta_{l_i,l_j}^{i,j})^2$ and

$$\| \tilde{\Theta} \|_F^2 = \sum_{i,j} \| \Theta_{ij} \|_F^2 = \sum_{i,j} \sum_{l_i,l_j} (\theta_{l_i,l_j}^{i,j})^2 .$$

The assertion has been demonstrated. □

LEMMA 5.50. *The estimate $\| \Theta \|_F \leq ck$ is valid, where c is independent of the level number k.*

Proof: As in Lemma 5.49, we introduce the block-matrix

$$\Theta_{k',k''} = \left[\theta_{ij}^{k',k''} \right]_{i,j} , \quad 1 \leq k', k'' \leq k,$$

and the matrix

$$\tilde{\Theta} = [\| \Theta_{k',k''} \|_F]_{k',k''=1}^{k} .$$

By Proposition 5.48, $\| \Theta_{k',k''} \|_F \leq c$. Computing the Frobenius norm of $\tilde{\Theta}$, one has

$$\| \tilde{\Theta} \|_F^2 \leq ck^2 .$$

By Lemma 5.49, one easily checks

$$\| \Theta \|_F = \| \tilde{\Theta} \|_F \leq ck$$

which is the desired result. □
The main result of this section is the upper eigenvalue estimate of the MTS-BPX preconditioner.

THEOREM 5.51. *For $u \in \mathbb{V}_k$, let*

$$|||u|||^2 = \min_{u = \sum_{l,i} u_i^l} \sum_{l=1}^{k} \sum_{i} a(u_i^l, u_i^l).$$

Then, one obtains

$$a(u,u) \leq ck \, ||| \, u \, |||^2 .$$

Proof: We give two proofs. The first proof follows by Lemma 3.1 and Lemma 3.5 of Zhang, [81], the fact $\| A \|_2 \leq \| A \|_F$ and Lemma 5.50.
In the second proof, we investigate the splitting $\mathbb{V}_k = \sum_{k'=1}^{k} \mathbb{U}_1^{k'} \oplus \mathbb{U}_2^{k'}$ (5.6.7). Now, let Θ be a $k \times k$ block matrix consisting of 2×2 matrices, i.e.

$$\Theta = \left[\theta^{k',k''}\right]_{k',k''=1}^{k} \quad \text{with} \quad \theta^{k',k''} = \left[\gamma_{\mathbb{U}_i^{k'},\mathbb{U}_j^{k''}}\right]_{i,j=1}^{2}.$$

By the usual Cauchy-inequality, the cosines $\gamma_{\mathbb{U}_i^{k'},\mathbb{U}_j^{k''}}$, cf. (5.6.25), of the angles between $\mathbb{U}_i^{k'}$ and $\mathbb{U}_j^{k''}$ are bounded from above by 1. Thus, $\| \theta^{k',k''} \|_F \leq 2$ follows. This is the analogous result of Proposition 5.48 for the space splitting (5.6.7). Using Lemma 5.49 and the proof of Lemma 5.50, the assertion follows. □

REMARK 5.52. *The eigenvalue estimate* $\lambda_{max}\left(\hat{C}_k^{-1} K_k\right) \leq ck$ *of the MTS-BPX preconditioner* $\hat{C}_k$ *for* K_k*, defined via relation (5.6.2), follows immediately.*

REMARK 5.53. *The constant in Theorem 5.51 depends linearly on the level number. The reason is the splitting into the spaces* $\mathbb{V}_i^l$*, not the differential operator. For the Laplacian, i.e.* $\omega(x) = 1$*, only this result can be proved using this space splitting.*

For this MTS-BPX preconditioner, Table 5.2 gives the lower and upper constants in the norm equivalence

$$\underline{c} \,|||\, u \,|||^2 \leq a(u,u) \leq \bar{c} \,|||\, u \,|||^2 \quad \forall u \in \mathbb{V}_l.$$

The constants are computed by a vector iteration and inverse vector iteration for the corresponding matrices in the case of the weight functions $\omega(\xi) = 1$ and $\omega(\xi) = \xi$. One can see that the constant $\bar{c}$ seems to be proportional to the level number for the weight functions $\omega(\xi) = 1$ and $\omega(\xi) = \xi$ which indicates that the estimate of Theorem 5.51 is sharp. The lower constant $\underline{c}$ seems to be bounded from below by a constant of about 0.488. However, we cannot prove the boundedness of $\underline{c}$ from below.

5.7 Implementational details

5.7.1 Fast solver for $C_{\mathbb{W}_k}$ and L_k

Using the Algorithm 5.3 $MULT$, linear systems of the type

$$C_{\mathbb{W}_l} \underline{u} = \underline{g}, \quad l = 2, \ldots, k, \tag{5.7.1}$$

have to be solved in order to apply the smoother $S_{0,l}$ (5.3.49), where $C_{\mathbb{W}_l}$ is defined via relation (5.3.48). Moreover, in order to apply the AMLI preconditioning system (5.4.5)

$$\underline{u} = \tilde{C}_{k,\mu,r}^{-1} \underline{g},$$

Level	$\underline{c}$		$\overline{c}$	
	$\omega(\xi)=1$	$\omega(\xi)=\xi$	$\omega(\xi)=1$	$\omega(\xi)=\xi$
2	0.607	0.748	1.86	1.78
3	0.522	0.647	2.73	2.59
4	0.495	0.583	3.44	3.39
5	0.489	0.543	4.00	4.03
6	0.488	0.524	4.45	4.52
7	0.488	0.512	4.81	4.91
8	0.488	0.504	5.11	5.23
9	0.488	0.498	5.35	5.60
10	0.488	0.495	5.55	6.11

Table 5.2: Lower and upper eigenvalue bounds of the MTS-BPX preconditioner.

systems of linear algebraic equations with the matrices $\tilde{C}_{22,l} = \left(1+\frac{2}{11}\sqrt{11}\right) C_{\mathbb{W}_l}$, $l=2,\ldots,k$, cf. (5.4.11), have to be solved. Therefore, it is important to find an efficient solution technique for the system (5.7.1). In this subsection, it will be shown that $C_{\mathbb{W}_k}$ is a block diagonal matrix consisting of tridiagonal blocks. Then, using Cholesky/Crout-decomposition, the system (5.7.1) can be solved in $\mathcal{O}(m_k)$ arithmetical operations, where m_k is the number of unknowns on level k, cf. subsection 2.2. Furthermore, we will show that the smoother $S_{0,k}$ (5.3.49) is a line smoother operating on lines ℓ_{2m-1} which will be defined below. According to (5.3.35), (5.3.39), and (5.3.46), the matrix $C_{\mathbb{W}_k}$ has the structure

$$C_{\mathbb{W}_k} = D_{\mathbb{W}_k} + R,$$

where $D_{\mathbb{W}_k}$ is the diagonal part of the matrix $K_{\mathbb{W}_k}$ defined in (5.3.47). The matrix R will be defined below. Let $b : \mathbb{W}_k \times \mathbb{W}_k \to \mathbb{R}$ be the following non-symmetric bilinear form uniquely determined by the values of the basis functions $\{\phi_{ij}^k\}_{(i,j)\in N_k} \in \mathbb{W}_k$

$$b(\phi_{ij}^k, \phi_{lm}^k) = \begin{cases} a(\phi_{ij}^k, \phi_{lm}^k) & \text{if} \quad \begin{array}{lll} i=l=2r-1, & j=2,\ldots,i, & m=j-1 \\ j=m=2r-1, & i=2,\ldots,j, & l=i-1 \end{array} \\ 0 & \text{otherwise} \end{cases}$$

for $r=1,\ldots,\frac{n}{2}$. By this definition, $a(\phi_{ij}^k, \phi_{lm}^k)$ is equal to the element $(i,j),(l,m)$ of the matrix K_k, if $i=l=2r-1$, $j=2,\ldots,i$, $m=j-1$, or $j=m=2r-1$, $i=2,\ldots,j$, $l=i-1$. The matrix R is defined as the symmetric part of the bilinear form b. More precisely, let

$$R = \left[b(\phi_{ij}^k, \phi_{lm}^k) + b(\phi_{lm}^k, \phi_{ij}^k)\right]_{(i,j),(l,m)\in N_k}.$$

After a proper permutation P, we have

$$C_{\mathbb{W}_k} = P^T \operatorname{blockdiag}\left[C_{\mathbb{W}_k,r}\right]_{r=0}^{\frac{n}{2}} P$$

with

$$C_{\mathbb{W}_k,r} = \begin{cases} \left[a(\phi_{ij}^k, \phi_{lm}^k)\right]_{(i,j),(l,m)\in\tilde{N}_{2r-1}} & \text{for} \quad r > 0 \\ \left[a(\phi_{ij}^k, \phi_{lm}^k)\right]_{(i,j),(l,m)\in\cup_{r=1}^{n/2-1}(\tilde{N}_{2r}\cap N_{k+1})} & \text{for} \quad r = 0 \end{cases}.$$

The index set $\tilde{N}_r$ is defined as

$$\tilde{N}_r = \left\{(i,j),(l,m) \in \{1,\ldots,r\}^4 : \; i = l = r \quad \text{or} \quad j = m = r\right\} \tag{5.7.2}$$

and N_k has been defined in (5.3.2). Thus, the matrices $C_{\mathbb{W}_k,r}$, $r \geq 1$, are tridiagonal matrices and the matrix $C_{\mathbb{W}_k,0}$ is a diagonal matrix. The shape functions of one block $C_{\mathbb{W}_k,r}$ correspond to one edge of the left picture of Figure 5.4 which is marked by a bold line. Therefore, the system (5.7.1) can be solved using Cholesky decomposition in $\mathcal{O}(n^2)$ flops. Hence, the operation $S_{0,k}\underline{w}$ is arithmetically optimal. Additionally, a smoother $S_{1,k} = I - \zeta L_k^{-1} K_k$ (5.5.3) has been built in

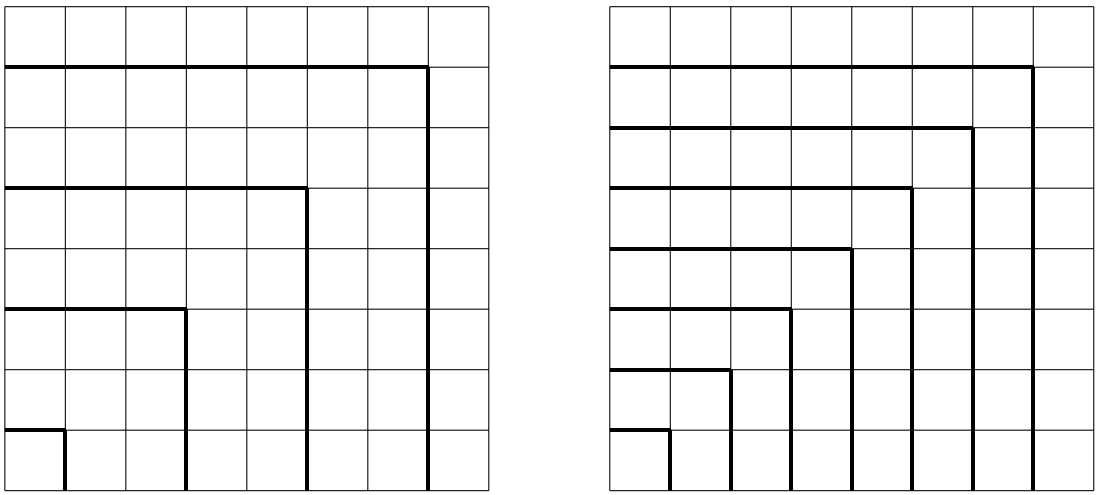

Figure 5.4: Nonzero entries of the matrices R (left) and $\tilde{R}$ (right).

subsection 5.5.1 which uses the ideas of $S_{0,k}$ (5.3.49). This smoother operates on the space $\mathbb{V}_k$. The matrix L_k can be interpreted as follows: Let

$$L_k = \operatorname{diag}(\mathfrak{l}) + \tilde{R},$$

where $\mathfrak{l} = \left[a(\phi_{ij}^k, \phi_{ij}^k)\right]_{(i,j)=(1,1)}^{(n-1,n-1)}$ and

$$\tilde{R} = \left[\tilde{b}(\phi_{ij}^k, \phi_{lm}^k) + \tilde{b}(\phi_{lm}^k, \phi_{ij}^k)\right]_{(i,j),(l,m)=(1,1)}^{(n-1,n-1)}$$

with the bilinear form $\tilde{b} : \mathbb{V}_k \times \mathbb{V}_k \to \mathbb{R}$,

$$\tilde{b}(\phi_{ij}^k, \phi_{lm}^k) = \begin{cases} a(\phi_{ij}^k, \phi_{lm}^k) & \text{if} \quad \begin{array}{ccc} i = l = r, & j = 2,\ldots,i, & m = j-1 \\ \text{or} \quad j = m = r, & i = 2,\ldots,j, & l = i-1 \end{array} \\ 0 & \text{otherwise} \end{cases} \tag{5.7.3}$$

for $r = 1, \ldots, n-1$. As well as $S_{0,k}$ (5.3.49), $S_{1,k}$ is a line smoother. However, it operates on each bold line in the right picture of Figure 5.4. So, we expect better convergence rates of a standard multi-grid algorithm (cf. Remark 5.4) in contrast to the smoother $S_{0,k}$. The matrix L_k is a block diagonal matrix consisting of tridiagonal blocks. After a proper permutation P,

$$L_k = P^T \text{blockdiag}\,[L_{k,r}]_{r=1}^{n-1}\, P,$$

where

$$L_{k,r} = \left[a(\phi_{ij}^k, \phi_{lm}^k)\right]_{(i,j),(l,m)\in\tilde{N}_r}$$

with the index set $\tilde{N}_r$ (5.7.2). The matrices $L_{k,r}$ are tridiagonal. The shape functions of one block $L_{k,r}$ correspond to nodes marked by one bold line in the right picture of Figure 5.4. Analogously to $S_{0,k}$, the operation

$$S_{1,k}\underline{w} = \underline{r}$$

can be done arithmetically optimal in $\mathcal{O}(n^2)$ flops using Cholesky- or Crout-decompostion. The same result is valid for the operation

$$\underline{w} = \hat{C}_k^{-1}\underline{r},$$

cf. relation (5.6.2).

5.7.2 Complexity of the algorithm

In this subsection, the arithmetical costs for the operations

$$u_{1,k} = MULT(k, u_{0,k}, g), \quad \text{cf. ALGORITHM 5.3} \tag{5.7.4}$$

using one of the smoothers $S_{0,l}$ (5.3.49), $S_{1,l}$ (5.5.3), $S_{2,l}$ (5.5.4), or $S_{3,l}$ (5.5.1) on level $l = 2, \ldots, k$ are considered. Moreover, it will be shown that the total cost for applying the AMLI preconditioner (5.4.5) and the MTS-BPX preconditioner (5.6.2), i.e.

$$\underline{w} = \tilde{C}_{k,r,\mu}^{-1}\underline{r}, \quad \text{(AMLI)} \tag{5.7.5}$$

$$\underline{w} = \hat{C}_k^{-1}\underline{r} \quad \text{(MTS-BPX)} \tag{5.7.6}$$

is arithmetically optimal.

THEOREM 5.54. *Let m_k be the number of unknowns on level k. Then, the arithmetical cost for each of the operations (5.7.4), (5.7.5), (5.7.6) is $\mathcal{O}(m_k)$, if the following assumptions are satisfied:*

- *$\mu \leq 3$ for (5.7.4) and (5.7.5),*
- *ν fixed for (5.7.4).*

Proof: At first, we consider the iteration (5.7.4). The number of arithmetical operations for (5.7.4) is denoted by $\mathfrak{W}_l$. By the definition of the parameter m_l,

$$m_l = \left(2^l - 1\right)^2$$

holds. The algorithm $MULT$ reads as follows:

1. pre-smoothing with ν pre-smoothing steps,
2. calculation and restriction of the defect,
3. solving the coarse grid system recursively for $\tilde{\mu} = 1, \ldots, \mu$,
4. interpolation and addition of the coarse grid correction,
5. post-smoothing with ν post-smoothing steps.

The cost of the step i on level l is denoted by $\mathfrak{W}_{i,l}$. Then, the cost for step 1 and the cost for step 5 can be estimated by

$$\mathfrak{W}_{1,l} = \mathfrak{W}_{5,l} \leq c_1 \nu m_l,$$

if one of the smoothers $S_{i,k}$, $i = 0, \ldots, 3$ is used, cf. subsection 5.7.1 for $S_{0,k}$ and $S_{1,k}$, see [40] for $S_{3,k}$ and see [42] for $S_{2,k}$. Since K_k (4.2.6) is a sparse matrix, one easily checks

$$\mathfrak{W}_{2,l} \leq c_2 m_l \qquad \text{and} \qquad \mathfrak{W}_{4,l} \leq c_4 m_l.$$

Moreover, by the definition of step 3, $\mathfrak{W}_{3,l} = \mu \mathfrak{W}_{l-1}$, $l \geq 2$. Then, by

$$\mathfrak{W}_l = \sum_{\zeta=1}^{5} \mathfrak{W}_{\zeta,l},$$

the recursive estimate

$$\mathfrak{W}_l \leq m_l(2\nu c_1 + c_2 + c_4) + \mu \mathfrak{W}_{l-1} \tag{5.7.7}$$

is valid. For $l = 2, \ldots, k$, the geometric series gives

$$\begin{aligned} \mathfrak{W}_k &\leq \sum_{l=2}^{k} m_l(2\nu c_1 + c_2 + c_4)\mu^{k-l} + \mathfrak{W}_1 \mu^{k-1} \\ &= (2\nu c_1 + c_2 + c_4) m_k \sum_{l=2}^{k} \mu^{k-l} \left(\frac{2^l - 1}{2^k - 1} \right)^2 + m_0 \mu^{k-1} \\ &\leq \nu c m_k, \end{aligned}$$

if $\mu < 4$. Therefore, the assertion has been established for the action (5.7.4). The remaining cases follow by the same arguments. □

5.8 Numerical examples

In this section, numerical experiments in order to solve (4.2.6), i.e.

$$K_k \underline{u}_k = \underline{g}_k, \tag{5.8.1}$$

are given. In subsection 5.8.1, the multi-grid algorithm in the version of the Algorithm 5.3 $MULT$ or Remark 5.4 is used as solution technique. In the following subsections, a preconditioned conjugate gradient method is the solver for (5.8.1). The preconditioners are the multi-grid preconditioner (5.5.8), cf. subsection 5.8.2, the AMLI preconditioner (5.4.5), cf. subsection 5.8.3, and the BPX preconditioners (5.6.2), (5.6.3), cf. subsection 5.8.4.

5.8.1 Convergence rates of multi-grid

In all experiments of this subsection, the multi-grid Algorithm 5.3 $MULT$ is used in order to solve (5.8.1). Written in vector form, the algorithm

$$\underline{u}_{j+1,k} = \underline{u}_{j,k} - (I - M_{k,S,\mu})K_k^{-1}(K_k\underline{u}_{j,k} - \underline{g}_k)$$

is used, where $M_{k,S,\mu}$ denotes the multi-grid operator with smoother S and the number of cycles μ. The following cases of initial values $u_{0,k}$ and right hand sides g in (5.1.1) are considered:

(A) $g = 0$ and $u_{0,k} = 1$,

(B) $g = 1$ and $u_{0,k} = 0$.

Using vectors of $\mathbb{R}^m$, the condition $g = 1$ means $\underline{g}_k = c\,[1,\ldots,1]^T$ (all triangles $\tau_{ij}^{u,k}$ of the triangulation have the same volume), and $u_{0,k} = 1$ means $\underline{u}_{0,k} = [1,\ldots,1]^T$. Two kinds of convergence rates are measured, the convergence rate ω_k in the Euclidian norm and the convergence rate σ_k, cf. (5.3.5), in the energy norm. More precisely, let

$$\begin{aligned}
\omega_k^2 &= \sup_{\underline{u}_{j,k}\neq\mathbf{0}} \frac{(\underline{u}_{j+1,k} - \underline{u}^*, \underline{u}_{j+1,k} - \underline{u}^*)}{(\underline{u}_{j,k} - \underline{u}^*, \underline{u}_{j,k} - \underline{u}^*)},\\
\sigma_k^2 &= \sup_{\underline{u}_{j,k}\neq\mathbf{0}} \frac{(K_k(\underline{u}_{j+1,k} - \underline{u}^*), \underline{u}_{j+1,k} - \underline{u}^*)}{(K_k(\underline{u}_{j,k} - \underline{u}^*), \underline{u}_{j,k} - \underline{u}^*)},
\end{aligned}$$

where $(\cdot,\cdot)$ is the Euclidian scalar product, and $\underline{u}^*$ denotes the exact solution of (5.8.1). In case (A), ω_k is measured, in case (B), the convergence rate σ_k in the energy norm is considered. Moreover, in all experiments, the algorithm is stopped, if the relative error in the Euclidian norm or in the energy-norm is less than $\varepsilon = 10^{-7}$. The upper tabular in Table 5.3 displays the numbers of iterations and the convergence rates σ_k of the multi-grid algorithm for (B) using the smoother $S_{0,k}$ (5.3.49) for $\mu = \mu_k = 1,\ldots,4$. The lower tabular in Table 5.3 shows the same results for ω_k in the case (A). The V-cycle ($\mu = 1$) has clearly growing numbers of iterations. For $\mu \geq 3$, we have mesh-independent convergence rates. It is not clear, if the rates of convergence σ_k are bounded from above by $\sigma < 1$ for the W-cycle ($\mu = 2$). The convergence rates σ_k do not depend on the choice of the right hand side and the initial value. More precisely, the maximal variance in the values of σ_k is 0.005 in all test examples considered for σ_k. Moreover, the number of smoothing steps $\nu \geq 1$ has no significant influence for the multi-grid convergence.
The convergence rates ω_k do not differ substantially from the convergence rates in the energy norm σ_k. For the smoother $S_{0,k}$, the rates are a slightly larger for $\mu \geq 2$ and lower for $\mu = 1$.

Level	$\mu = 1$		$\mu = 2$		$\mu = 3$		$\mu = 4$	
	It	σ_k	It	σ_k	It	σ_k	It	σ_k
2	18	0.4070	18	0.4070	18	0.4070	18	0.4070
3	32	0.6017	24	0.4997	22	0.4778	22	0.4722
4	50	0.7239	25	0.5221	22	0.4698	21	0.4583
5	72	0.7974	27	0.5449	22	0.4770	21	0.4582
6	97	0.8463	30	0.5755	24	0.5035	22	0.4719
7	128	0.8814	34	0.6201	25	0.5156	22	0.4788
8	176	0.9123	37	0.6432	26	0.5282	23	0.4838
9	247	0.9373	41	0.6724	26	0.5339	23	0.4847
10	346	0.9545	44	0.6901	26	0.5380	23	0.4841

Level	$\mu = 1$		$\mu = 2$		$\mu = 3$		$\mu = 4$	
	It	ω_k	It	ω_k	It	ω_k	It	ω_k
2	18	0.4013	18	0.4013	18	0.4013	18	0.4013
3	30	0.5779	21	0.4621	20	0.4409	20	0.4359
4	45	0.6946	22	0.4709	20	0.4463	20	0.4462
5	60	0.7611	27	0.5399	22	0.4775	22	0.4711
6	74	0.8040	30	0.5806	25	0.5156	23	0.4914
7	93	0.8409	35	0.6253	26	0.5370	24	0.5078
8	127	0.8800	39	0.6583	28	0.5550	25	0.5200
9	171	0.9098	43	0.6852	29	0.5690	26	0.5294
10	235	0.9336	48	0.7105	30	0.5803	26	0.5371

Table 5.3: Mg-convergence rates ω_k (below) and σ_k (above) using smoother $S_{0,k}$ ($\nu = 1$).

For the V-cycle, the results are not satisfactory. The reason for the bad convergence of the V-cycle is the smoother $S_{0,k}$ which operates on the nodes corresponding to the space $\mathbb{W}_k$ only.

In subsection 5.5.1, cf. relations (5.5.3), (5.5.4), (5.5.1), smoothers $S_{i,k}$, $i = 1, 2, 3$ are defined which work on the space $\mathbb{V}_k$. The multi-grid Algorithm 5.3 in the version of Remark 5.4 shows mesh-size independent convergence rates $\sigma_k < \sigma < 1$ for the V-cycle using these smoothers, cf. Table 5.4 for case (B). For the smoothers $S_{1,k}$ and $S_{2,k}$, the parameter $\omega = 0.8$ is chosen in relations (5.5.3) and (5.5.4). This relaxation parameter shows the best mg-convergence rates σ_k. For the W-cycle, the convergence rates using these smoothers do not change significantelly from that of the V-cycle. We refer to the preprints [13], [11] for more numerical examples.

Now, we compare all these smoothers. On the left picture of Figure 5.5, the multi-grid convergence rates σ_k for all smoothers are compared. The time measured in seconds which is needed in order to reduce the relative error in the energy norm up to a factor of $\varepsilon = 10^{-7}$, is displayed on the right picture. For a better visibility, the time is scaled with the number of unknowns.

It can be concluded from the results that the ILU-smoother $S_{2,k}$ (5.5.4) and the line Gauß-Seidel smoother $S_{3,k}$ (5.5.1) are the best smoothers. Moreover, the mg-algorithm using these smoothers are the fastest ones. The smoother for which Theorem 5.36 holds, the smoother $S_{0,k}$ with $\mu = 3$,

Level	$S_{1,k}$		$S_{2,k}$		$S_{3,k}$	
	It	σ_k	It	σ_k	It	σ_k
2	9	0.1611	6	0.0614	3	0.0014
3	11	0.2290	8	0.1007	5	0.0234
4	13	0.2723	8	0.1224	6	0.0512
5	15	0.3250	9	0.1348	6	0.0639
6	16	0.3517	9	0.1399	7	0.0705
7	16	0.3619	9	0.1421	7	0.0780
8	17	0.3680	9	0.1434	7	0.0853
9	17	0.3720	9	0.1447	7	0.0912
10	17	0.3750	9	0.1470	7	0.0960

Table 5.4: Convergence rates σ_k of multi-grid algorithm $MULT$ using smoothers $S_{i,k}$, $i = 1, 2, 3$ with $\nu = 1$.

presents high convergence rates and is relatively expensive.
Now, consider the convergence rates ω_k. We will see that these rates can depend on the special choice of the initial value. The convergence rate is given by the spectral radius of the mg-operator $M_{k,\mu,S}$. Usually, the vector $\underline{u}_{0,k}$ has a non-vanishing component in the eigenbasis of $M_{k,S,\mu}$ to that eigenvector which corresponds to the dominant eigenvalue. However, it can be possible that we have chosen an $\underline{u}_{0,k}$ with zero component to that eigenvector. For this reason, several examples are considered. In all examples, we set $g = 0$, whereas the initial value $u_{0,k}$ is chosen as follows:

(a) $\underline{u}_{0,k} = \mathfrak{e}$, σ_k is considered instead of ω_k,

(b) $\underline{u}_{0,k} = \underline{v}_k \otimes \underline{w}_k$, where v_k and w_k are chosen randomly,

(c) $\underline{u}_{0,k} = \mathfrak{e}$,

(d) $\underline{u}_{0,k} = \underline{v}_k \otimes \underline{w}_k$, where $v_k = \mathfrak{e}$ and w_k is chosen randomly,

(e) $\underline{u}_{0,k} = \underline{v}_k \otimes \underline{w}_k$, where $v_k = \left[\sin \frac{3i}{n-1}\right]_{i=1}^{n-1}$ and $w_k = \left[\cos \frac{i}{n-1}\right]_{i=1}^{n-1}$,

(f) $\underline{u}_{0,k} = \underline{v}_k + \underline{w}_k$, where $v_k = \left[\sin \frac{3i}{n-1}\right]_{i=1}^{n-1} \otimes \mathfrak{e}$ and $w_k = \mathfrak{e} \otimes \left[\cos \frac{i}{n-1}\right]_{i=1}^{n-1}$,

with

$$\mathfrak{e} = [1, \ldots, 1]^T .$$

The following smoothers for the multi-grid algorithm are considered:

(i) smoother $S_{3,k}$ (5.5.1) with $\mu = 1$,

(ii) ILU-smoother $S_{2,k}$ (5.5.4) with $\mu = 1$ and $\omega = 0.8$,

(iii) smoother $S_{1,k}$ (5.5.3) with $\mu = 1$ and $\omega = 0.8$,

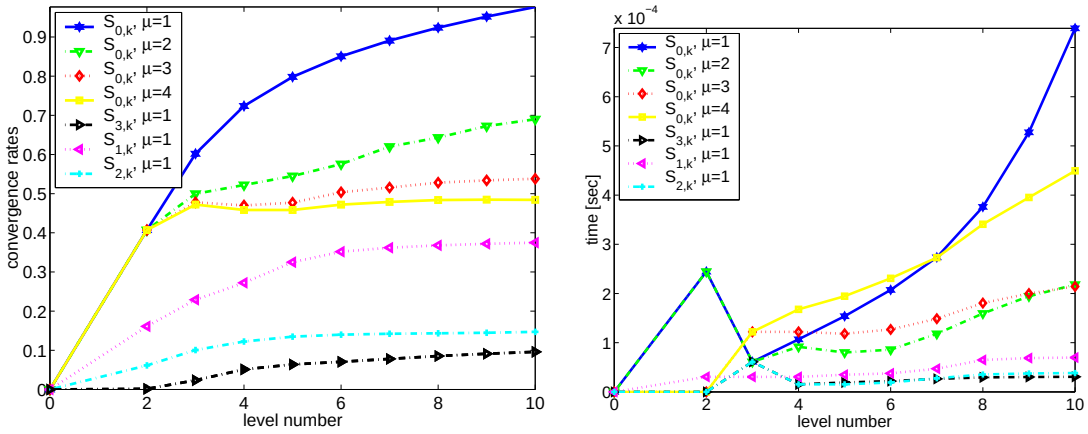

Figure 5.5: Comparison of all smoothers, mg-convergence (left), $\frac{t}{4^k}$ (right), where $t \ldots$ time in seconds, $k \ldots$ level number.

(iv) smoother $S_{0,k}$ (5.3.49) with $\mu = 3$.

The convergence rates ω_k are displayed in Figure 5.6.
One can see that the convergence rates ω_k depend on the choice of the initial value. In case (iii) only, the convergence rate ω_k is lower than σ_k. The examples (c), (e) and (f) show nearly the same convergence rates in all cases. For (ii) and (iv), the l_2-mg-convergence rate ω_k is slightly higher than the energetic multi-grid convergence rate σ_k.

5.8.2 Multi-grid preconditioner

In the following three subsections, the preconditioned conjugate gradient method is used as solver for (5.8.1). In this subsection, cf. subsection 5.5.2, the preconditioner (5.5.8), i.e.

$$\bar{C}^{-1}_{k,S,\mu} = (I - M_{k,\mu,S})K_k^{-1},$$

is used. In all experiments, $\underline{g}_k = [1, \ldots, 1]^T$ is chosen as right hand side of (5.8.1). The algorithm is stopped, if the relative error in the preconditioned energy norm is reduced up to a factor of 10^{-9}. Table 5.5 displays the number of iterations of the pcg-method using the smoothers $S_{0,k}$ (5.3.49), $S_{1,k}$ (5.5.3) with $\omega = 0.8$, $S_{2,k}$ (5.5.4) with $\omega = 0.8$, $S_{3,k}$ (5.5.1) and the Gauß-Seidel (GS) smoother.
For $S_{0,k}$ with $\mu = 1$, there is a logarithmic growth of the number of iterations. The multi-grid preconditioner with the Gauß-Seidel smoother (GS) shows clearly growing number of iterations. In all other cases, the results indicate the boundedness of the numbers of iterations by some small constant.

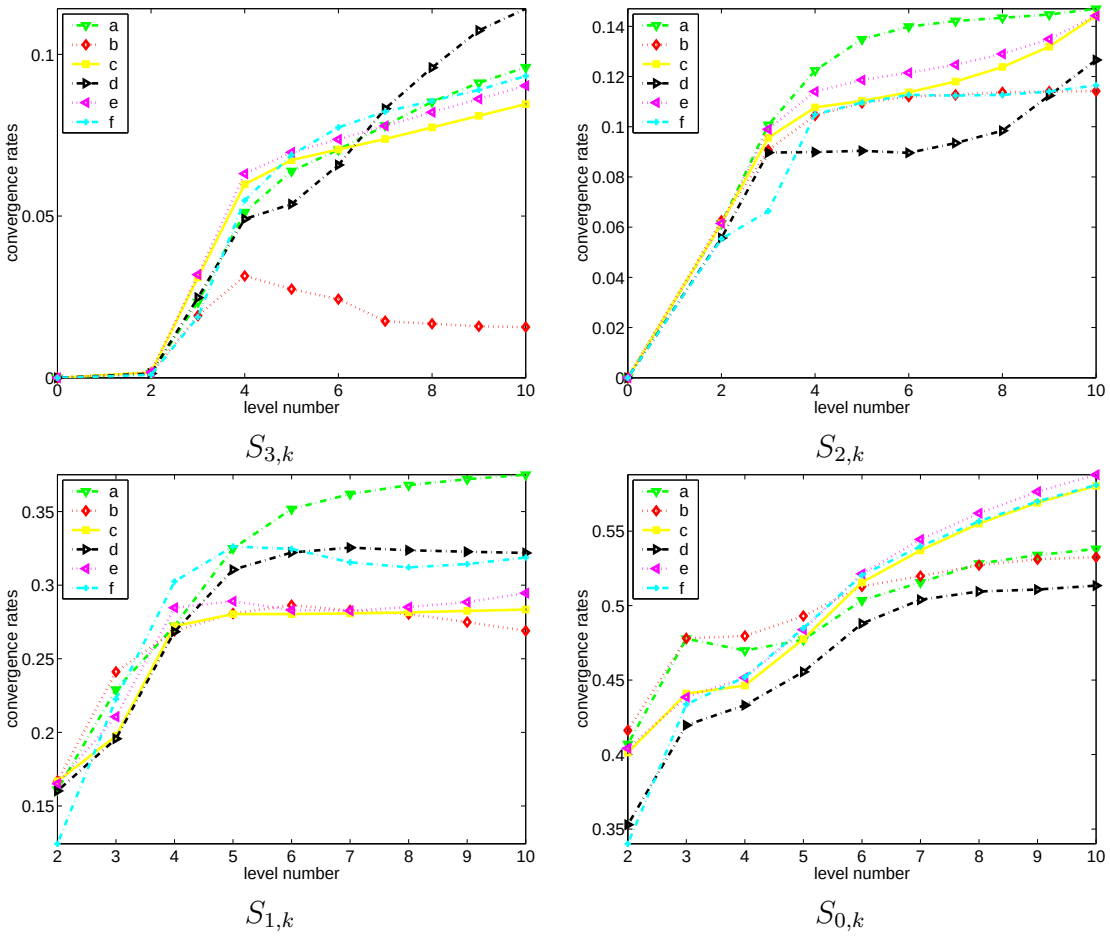

Figure 5.6: Comparison of ω_k for several examples.

5.8.3 AMLI preconditioner

Consider (5.8.1) and solve this linear system with the preconditioned conjugate gradient method. The algorithm is stopped, if the relative error measured in the preconditioned energy norm is lower than $\varepsilon = 10^{-9}$. The right hand side $\underline{g}_k = [1, \ldots, 1]^T$ is chosen. Now, the AMLI preconditioner $\tilde{C}_{k,r,\mu}$ (5.4.5) is used as preconditioner for K_k with the polynomials $P_{\mu,r}(t)$

$$\begin{array}{rcll} P_{\mu,1}(t) & = & (1-t)^{\mu} & \text{for} \quad \mu = 1, 2, 3, \\ P_{2,r}(t) & = & (1-rt)^2 & \text{for} \quad r = \frac{52}{35}, \frac{17}{9} \end{array} \tag{5.8.2}$$

and the matrix $\tilde{C}_{22,l}$ defined in relation (5.4.11). Note that Theorem 5.39 holds for $P_{\mu,r}(t) = P_{2,\frac{17}{9}}(t)$. Table 5.6 displays the number of iterations for the AMLI preconditioners with the polynomials (5.8.2). Recall that for the definition of the polynomial $P_{\mu,r}(t)$ of the AMLI preconditioner (5.4.5), the eigenvalue bounds $\lambda_{min}\left(\mathbb{C}_{\mathbb{W}_k}{}^{-1} K_{\mathbb{W}_k}\right)$ and $\lambda_{max}\left(\mathbb{C}_{\mathbb{W}_k}{}^{-1} K_{\mathbb{W}_k}\right)$ are required, see (5.4.9). However, the eigenvalue bounds $\lambda_{min}\left(\mathbb{C}_{\mathbb{W}_k}{}^{-1} K_{\mathbb{W}_k}\right)$ and $\lambda_{max}\left(\mathbb{C}_{\mathbb{W}_k}{}^{-1} K_{\mathbb{W}_k}\right)$ in Theorem 5.33 are estimates and the exact values are not known. Probably, the exact values can be

Level	$S_{0,k}$			$S_{1,k}$	$S_{2,k}$	$S_{3,k}$	GS
	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 1$	$\mu = 1$	$\mu = 1$	$\mu = 1$
2	7	8	7	7	6	5	8
3	12	12	11	9	7	6	11
4	15	13	13	10	7	6	13
5	16	14	13	10	7	6	18
6	18	14	13	11	7	6	25
7	21	15	13	11	7	7	33
8	23	16	14	11	7	7	44
9	25	16	14	11	7	7	59

Table 5.5: Number of iterations of the pcg-method using a multi-grid preconditioner $M_k^{\mu,S}$ with $S = S_{i,k}$, $i = 0, \ldots, 3$.

Level	$P_{1,1}(t)$	$P_{2,1}(t)$	$P_{2,\frac{52}{35}}(t)$	$P_{2,\frac{17}{9}}(t)$	$P_{3,1}(t)$	mixed
2	8	8	8	8	8	8
3	17	16	16	16	16	18
4	23	17	17	18	17	21
5	28	18	17	19	18	23
6	33	19	17	21	18	24
7	39	20	18	21	18	25
8	46	21	18	21	18	26
9	52	22	17	22	18	26

Table 5.6: Number of iterations of the pcg-method with AMLI preconditioners.

better. The polynomial $P_{2,\frac{52}{35}}(t)$ is that polynomial $(1 - rt)^2$ with the smallest number of iterations on level 9 for $r = \frac{36}{35}, \frac{38}{35}, \ldots, \frac{66}{35}$. Furthermore, we used the AMLI preconditioner on level k using the polynomial $1 - t$ on the levels $l = 1, 3, \ldots$ and the polynomial $(1 - \frac{17}{9}t)^2$ on the levels $l = 2, 4, \ldots$, where $l \leq k$. This case is denoted by mixed in the last column of Table 5.6.
The number of iterations are bounded by a constant for $P_{2,\frac{52}{35}}(t)$, $P_{2,\frac{17}{9}}(t)$, $P_{3,1}(t)$ and for the case of $P_{1,1}(t)$ on each odd and $P_{2,\frac{17}{9}}(t)$ on each even level. However, they grow proportional to the number of levels for $P_{1,1}(t)$ and $P_{2,1}(t)$.

5.8.4 BPX preconditioner

Finally, numerical results are given in order to solve (5.8.1) with the pcg-method and the MTS-BPX preconditioner $\hat{C}_k$ (5.6.2) and the ILU-BPX preconditioner $\hat{\mathcal{C}}_k$ (5.6.3). As before, $\underline{g}_k = [1, \ldots, 1]^T$ is chosen. Table 5.7 displays the number of iterations for several relative accuracies ε in the preconditioned energy norm. The results are compared with the results of the BPX pre-

Level	MDS-BPX	MTS-BPX			ILU-BPX
	$\varepsilon = 10^{-5}$	$\varepsilon = 10^{-5}$	$\varepsilon = 10^{-9}$	$\varepsilon = 10^{-16}$	$\varepsilon = 10^{-9}$
2	9	8	9	9	8
3	16	11	18	27	14
4	24	14	23	37	19
5	33	15	26	44	21
6	44	16	28	49	23
7	58	17	30	52	24
8	76	17	31	56	25
9	97	18	32	58	26

Table 5.7: Number of iterations of the PCG-method in order to solve $K_p \underline{u}_p = \underline{f}_p$ with the preconditioners $\hat{C}_k$ and $\hat{\mathcal{C}}_k$.

conditioner with multiple diagonal scaling (MDS), see [10]. The multiple tridiagonal scaling procedure and the ILU-decomposition stabilizes the BPX preconditioner. The number of iterations grow moderately. In comparison to the multi-grid preconditioners of subsection 5.8.2, the numbers of iterations are larger. However, the solution of a preconditioned system with a BPX-like preconditioner is cheaper than the solution with a multi-grid preconditioner.

6 Multi-level preconditioner for p-fem

In this chapter, we return to the p-version of the fem. The linear system of algebraic finite element equations

$$A_{\mathcal{R}_2}\underline{u}_p = \underline{f}_p \tag{6.1}$$

with the matrix $A_{\mathcal{R}_2}$ (3.3.4) is considered.

6.1 Final estimates of the condition numbers

We are interested in a good preconditioner for the matrix $A_{\mathcal{R}_2}$ (3.3.4), the element stiffness matrix for the interior unknowns on $\mathcal{R}_2 = (-1,1)^2$ with respect to the basis of the integrated Legendre polynomials $\{\hat{L}_{ij}\}_{i,j=2}^{p}$. Two preconditioners will be introduced. Let P be the permutation of Proposition 3.3, $\tilde{C}_{k,r,\mu}$ the AMLI preconditioner (5.4.5) with the polynomial $P_{\mu,r}(t)$ and $\bar{C}_{k,S,\mu}$ be the multi-grid preconditioner (5.5.8) with smoother S. Via these matrices, the multi-level preconditioners

$$\tilde{\mathfrak{M}}_{k,r,\mu} = P^T \text{blockdiag}\left[2n^2\tilde{C}_{k,r,\mu}\right]_{i=1}^{4} P, \tag{6.1.1}$$

$$\bar{\mathfrak{M}}_{k,S,\mu} = P^T \text{blockdiag}\left[2n^2\bar{C}_{k,S,\mu}\right]_{i=1}^{4} P \tag{6.1.2}$$

are defined, where k denotes the level number, $n = 2^k$ and $p = 2n-1$ is the polynomial degree.

THEOREM 6.1. *The eigenvalue estimates*

$$\lambda_{min}\left(\tilde{\mathfrak{M}}_{k,r,\mu}^{-1}A_{\mathcal{R}_2}\right) \asymp 1, \qquad \lambda_{max}\left(\tilde{\mathfrak{M}}_{k,r,\mu}^{-1}A_{\mathcal{R}_2}\right) \preceq 1+\log p, \tag{6.1.3}$$

$$\lambda_{min}\left(\bar{\mathfrak{M}}_{k,S,\mu}^{-1}A_{\mathcal{R}_2}\right) \asymp 1, \qquad \lambda_{max}\left(\bar{\mathfrak{M}}_{k,S,\mu}^{-1}A_{\mathcal{R}_2}\right) \preceq 1+\log p \tag{6.1.4}$$

are valid for the polynomial $P_{\mu,r}(t) = \left(1-\frac{17}{9}t\right)^2$ and the matrix $\tilde{C}_{22,k}$ (5.4.11) in (6.1.3), and for $\mu = 3$ and the smoother $S = S_{0,k}$ (5.3.49) in (6.1.4).

Proof: By Theorem 5.41, we have

$$\bar{C}_{k,S,\mu} \asymp K_k$$

for $\mu = 3$ and $S = S_{0,k}$ defined in (5.3.49). By Theorem 5.39, we have

$$\tilde{C}_{k,r,\mu} \asymp K_k$$

with the parameters $\mu = 2$ and $r = \frac{17}{9}$ in (5.4.13), where K_k is the matrix (4.2.6). Furthermore by Lemma 4.3, $K_k = \frac{1}{2n^2} C_4$ follows, cf. (3.4.18). Hence, one can deduce

$$\begin{aligned}
\lambda_{min}\left((2n^2\bar{C}_{k,S,\mu})^{-1}C_4\right) &\asymp 1,\\
\lambda_{max}\left((2n^2\bar{C}_{k,S,\mu})^{-1}C_4\right) &\asymp 1,\\
\lambda_{min}\left((2n^2\tilde{C}_{k,r,\mu})^{-1}C_4\right) &\asymp 1,\\
\lambda_{max}\left((2n^2\tilde{C}_{k,r,\mu})^{-1}C_4\right) &\asymp 1.
\end{aligned}$$

Using Theorem 3.11 with $\lambda_{min}\left({C_4}^{-1}\mathfrak{A}_1\right) \asymp 1$ and $\lambda_{max}\left({C_4}^{-1}\mathfrak{A}_1\right) \preceq 1 + \log n$, and Proposition 3.3 with $\mathfrak{A}_i \asymp \mathfrak{A}_1$ for $i = 2, 3, 4$, one can conclude that

$$\begin{aligned}
\lambda_{min}\left((2n^2\bar{C}_{k,S,\mu})^{-1}\mathfrak{A}_i\right) &\asymp 1,\\
\lambda_{max}\left((2n^2\bar{C}_{k,S,\mu})^{-1}\mathfrak{A}_i\right) &\preceq 1 + \log p,\\
\lambda_{min}\left((2n^2\tilde{C}_{k,r,\mu})^{-1}\mathfrak{A}_i\right) &\asymp 1,\\
\lambda_{max}\left((2n^2\tilde{C}_{k,r,\mu})^{-1}\mathfrak{A}_i\right) &\preceq 1 + \log p,
\end{aligned}$$

where $n = \frac{p+1}{2}$. By the first assertion of Proposition 3.3, i.e.

$$A_{\mathcal{R}_2} = P^T \text{blockdiag}\left[\mathfrak{A}_i\right]_{i=1}^4 P$$

holds with some permutation P, the assertions follow immediately.□
Thus, we have found two nearly asymptotically optimal methods in order to solve the system of linear algebraic equations (6.1).

6.2 Numerical results

In this subsection, numerical results in order to solve

$$A_{\mathcal{R}_2}\underline{u}_p = \underline{f}_p \tag{6.2.1}$$

using the preconditioned conjugate gradient method are given. In all experiments, the right hand side

$$\underline{f}_p = \left[\begin{array}{cccc} 1 & 1 & \dots & 1 \end{array}\right]^T$$

is chosen. The algorithm is stopped, if the relative error in the preconditioned energy norm is reduced up to the factor $\varepsilon = 10^{-9}$. All calculations are done on a Pentium-III, 800 MHz.

p	$\bar{\mathfrak{M}}_{k,S_{1,k},1}$		$\bar{\mathfrak{M}}_{k,S_{0,k},1}$		$\bar{\mathfrak{M}}_{k,S_{0,k},2}$		$\bar{\mathfrak{M}}_{k,S_{0,k},3}$	
	It	time [sec]	It	time [sec]	It	time [sec]	It	time [sec]
7	15	0.004	16	0.004	16	0.004	16	0.004
15	17	0.015	20	0.015	20	0.023	20	0.031
31	20	0.059	26	0.074	23	0.094	23	0.141
63	21	0.250	31	0.352	24	0.371	24	0.578
127	22	1.21	36	1.87	26	1.78	25	2.53
255	23	6.08	42	10.5	28	8.37	26	11.5
511	24	31.5	50	61.0	29	41.6	27	55.1
1023	24	133.	59	303.	30	186.	28	249.

Table 6.1: Numbers of iterations of the pcg-method for $A_{\mathcal{R}_2}$ using several multi-grid preconditioners $\bar{\mathfrak{M}}_{k,S,\mu}$.

6.2.1 Multi-grid preconditioner

Table 6.1 displays the numbers of iterations and the time to reduce the error using the preconditioner $\bar{\mathfrak{M}}_{k,S,\mu}$ with $S = S_{0,k}$ defined in (5.3.49) for $\mu = 1, 2, 3$, and $S = S_{1,k}$ defined in (5.5.3) for $\mu = 1$. In the two cases $\bar{\mathfrak{M}}_{k,S_{1,k},1}$ and $\bar{\mathfrak{M}}_{k,S_{0,k},3}$, the numbers of iterations grow slightly. For $\bar{\mathfrak{M}}_{k,S_{0,k},1}$, there is a stronger increase of the numbers of iterations. The preconditioner $\bar{\mathfrak{M}}_{k,S_{0,k},3}$, for which Theorem 6.1 holds, is relatively slow in reducing the error in comparison to all other preconditioners. For example, $\bar{\mathfrak{M}}_{k,S_{0,k},1}$ is faster for $p \leq 255$, although the numbers of iterations grow relatively fast.
However, the numbers of iterations are not bounded by a constant independent of p in all experiments. Next as preconditioner for $\mathfrak{A}_1$, we consider the multi-grid preconditioner $\breve{C}_{k,S,\mu}$ (5.5.12) arising from the discretization of (4.2.2), cf. Remark 4.2, i.e.

$$\begin{aligned} -2\left(y^2 u_{xx} + x^2 u_{yy}\right) + \left(\frac{x^2}{y^2} + \frac{y^2}{x^2}\right) u &= g \quad \text{in } \Omega = (0,1)^2, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

The corresponding system of linear algebraic equations of the finite difference discretization of this problem, cf. subsection 5.5.3, can be solved by a multi-grid algorithm using a smoother S. Let

$$\breve{\mathfrak{M}}_{k,S,\mu} = P^T \text{blockdiag}\left[\breve{C}_{k,S,\mu}\right]_{i=1}^{4} P \tag{6.2.2}$$

be the corresponding multi-grid preconditioner, where P denotes the permutation of Proposition 3.3. Table 6.2 displays the numbers of iterations for the mg-preconditioners $\bar{\mathfrak{M}}_{k,S,\mu}$ and $\breve{\mathfrak{M}}_{k,S,\mu}$ with $\mu = 1$ and the smoothers $S = S_{2,k}$ defined in (5.5.4) and $S = S_{3,k}$ defined in (5.5.1) for K_k (4.2.6) and C_1 (3.4.3), respectively. One can see that, in contrast to the mg-convergence rates considered in section 5.8, the choice of the smoothers $S = S_{1,k}$, cf. Table 6.1, $S = S_{2,k}$, or $S = S_{3,k}$ for the preconditioner $\bar{\mathfrak{M}}_{k,S,1}$ does not influence the results so significantelly. However,

p	$\bar{\mathfrak{M}}_{k,S_{2,k},1}$		$\breve{\mathfrak{M}}_{k,S_{2,k},1}$		$\bar{\mathfrak{M}}_{k,S_{3,k},1}$		$\breve{\mathfrak{M}}_{k,S_{3,k},1}$	
	It	time [sec]	It	time [sec]	It	time [sec]	It	time [sec]
7	15	0.004	16	0.004	15	0.008	16	0.008
15	17	0.019	16	0.019	17	0.023	16	0.027
31	19	0.062	16	0.062	20	0.105	16	0.105
63	21	0.269	16	0.254	21	0.461	16	0.453
127	22	1.30	16	1.16	21	1.99	16	1.94
255	23	7.03	16	5.77	22	9.19	16	8.43
511	23	34.8	16	26.6	23	41.3	16	35.9
1023	23	147.	16	111.	23	168.	16	146.

Table 6.2: Numbers of iterations of the pcg-method for $A_{\mathcal{R}_2}$ using several multi-grid preconditioners $\bar{\mathfrak{M}}_{k,S,1}$ and $\breve{\mathfrak{M}}_{k,S,1}$ and $S = S_{2,k}$, $S = S_{3,k}$.

from the preconditioners $\bar{\mathfrak{M}}_{k,S,1}$ with the three smoothers $S_{1,k}$, $S_{2,k}$, and $S_{3,k}$, the preconditioner with $S = S_{1,k}$ is cheaper than the other ones and the fastest. The best multi-grid preconditioner are the preconditioners $\breve{\mathfrak{M}}_{k,S,\mu}$ which indicate constant numbers of iterations.

6.2.2 AMLI preconditioner

p	$\tilde{\mathfrak{M}}_{k,1,1}$		$\tilde{\mathfrak{M}}_{k,2,1}$		$\tilde{\mathfrak{M}}_{k,1,\frac{17}{9}}$		$\tilde{\mathfrak{M}}_{k,1,\frac{12}{7}}$	
	It	time [sec]	It	time [sec]	It	time [sec]	It	time [sec]
7	16	0.008	16	0.004	18	0.008	17	0.008
15	22	0.016	22	0.031	23	0.023	22	0.023
31	28	0.101	25	0.125	26	0.133	26	0.125
63	34	0.531	28	0.602	29	0.617	28	0.593
127	43	3.27	31	3.09	31	3.04	29	3.86
255	51	20.6	33	16.2	33	16.0	30	14.6
511	61	130.	35	87.8	34	84.3	31	77.0
1023	73	671.	37	411.	34	375.	31	342.

Table 6.3: Numbers of iterations of the pcg-method for $A_{\mathcal{R}_2}$ using several AMLI preconditioners $\tilde{\mathfrak{M}}_{k,r,\mu}$.

In this subsection, the system (6.1) is solved by the pcg-method with the AMLI preconditioner $\tilde{\mathfrak{M}}_{k,r,\mu}$ (6.1.1). Table 6.3 displays the numbers of iterations and time to reduce the error in the preconditioned energy norm up to a factor 10^{-9} using the polynomial iteration

$$P_{\mu,r}(t) = (1 - rt)^{\mu}.$$

A slight increase of the numbers of iterations can be seen in the two cases $P(t) = (1 - \frac{12}{7}t)^2$ and $P(t) = (1 - \frac{17}{9}t)^2$. For $P(t) = (1 - t)$, similar to the V-cycle of multi-grid, there is a stronger growth of the numbers of iterations. The method using the preconditioner $\tilde{\mathfrak{M}}_{k,\frac{12}{7},2}$, in which the polynomial $P(t) = (1 - \frac{12}{7}t)^2$ is used, is the fastest AMLI preconditioner.
However, the comparison of the results for the AMLI preconditioners of Table 6.3 with the multi-grid preconditioners of Table 6.2 and Table 6.1 shows substantially lower numbers of iterations for most multi-grid preconditioners than for each of the AMLI preconditioners. Moreover, less time is needed in order to reduce the error.
If we compare the preconditioners $\bar{\mathfrak{M}}_{k,S_{0,k},3}$ and $\tilde{\mathfrak{M}}_{k,\frac{17}{9},2}$ of Theorem 6.1, the numbers of iterations are lower for $\bar{\mathfrak{M}}_{k,S_{0,k},3}$. However, solving (6.1) using the preconditioner $\bar{\mathfrak{M}}_{k,S_{0,k},3}$ requires about two third of the time in order to reduce the error up to a factor of 10^{-9} of the time needed using the preconditioner $\tilde{\mathfrak{M}}_{k,\frac{17}{9},2}$.

6.2.3 BPX preconditioner

In this subsection, the MTS-BPX preconditioner $\hat{C}_k$ (5.6.2) or the ILU-BPX preconditioner $\hat{\mathcal{C}}_k$ (5.6.3) is considered on each block $\mathfrak{A}_1$. Via the permutation matrix P of relation (6.1.1), the preconditioner

$$\hat{\mathfrak{M}}_k = P^T \text{blockdiag} \left[2n^2 \hat{C}_k \right]_{i=1}^{4} P \tag{6.2.3}$$

is introduced. If we replace $\hat{C}_k$ by $\hat{\mathcal{C}}_k$ in (6.2.3), the preconditioner $\hat{\mathfrak{M}}_{k,ILU}$ is defined. Table 6.4 displays the numbers of iterations and the time to reduce the error up to a factor of $\varepsilon = 10^{-9}$ in order to solve (6.1) using $\hat{\mathfrak{M}}_k$, or $\hat{\mathfrak{M}}_{k,ILU}$ as preconditioner. The numbers of iterations grow

p	MTS-BPX		ILU-BPX
	It	time [sec]	It
7	17	0.004	17
15	24	0.008	24
31	28	0.039	28
63	32	0.195	32
127	37	1.18	36
255	42	6.41	40
511	46	31.6	44
1023	50	141.	47

Table 6.4: Numbers of iterations of the pcg-method with preconditioner $\hat{\mathfrak{M}}_k$ and $\hat{\mathfrak{M}}_{k,ILU}$.

as $1 + \log p$. In comparison to the multi-grid preconditioners $\breve{\mathfrak{M}}_{k,S,\mu}$ (6.2.2) and $\bar{\mathfrak{M}}_{k,S,\mu}$ (6.1.2), the preconditioners $\hat{\mathfrak{M}}_k$ and $\hat{\mathfrak{M}}_{k,ILU}$ show relatively large numbers of iterations. However, the time in order to reduce the error is about the same. In the next subsection, a more pro-founding comparison is given.

6.2.4 Comparison of all preconditioners

In this subsection, the time is measured which is required to reduce the error up to a factor of $\varepsilon = 10^{-9}$ in order to solve the linear system (6.1). The results are displayed in Figure 6.1. For reasons of a better visibility, all results are scaled with p^2, where p is the polynomial degree. For the time, a logarithmic scaling is used. The following preconditioners are considered:

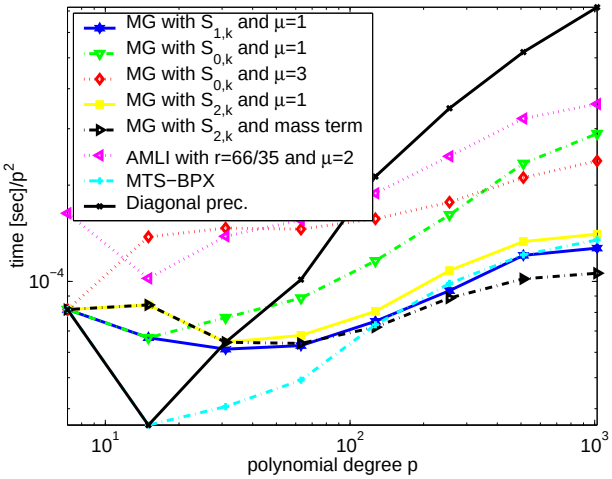

Figure 6.1: Comparison of several preconditioners.

- the multi-grid preconditioner $\bar{\mathfrak{M}}_{k,S_{1,k},1}$, denoted by MG with $S_{1,k}$ and $\mu = 1$,
- the multi-grid preconditioner $\bar{\mathfrak{M}}_{k,S_{0,k},1}$, denoted by MG with $S_{0,k}$ and $\mu = 1$,
- the multi-grid preconditioner $\bar{\mathfrak{M}}_{k,S_{0,k},3}$, denoted by MG with $S_{0,k}$ and $\mu = 3$,
- the multi-grid preconditioner $\bar{\mathfrak{M}}_{k,S_{2,k},1}$, denoted by MG with $S_{2,k}$ and $\mu = 1$,
- the multi-grid preconditioner $\breve{\mathfrak{M}}_{k,S_{2,k},1}$, denoted by MG with $S_{2,k}$ and mass term,
- the AMLI preconditioner $\tilde{\mathfrak{M}}_{k,\frac{17}{9},2}$, denoted by AMLI with $r = \frac{17}{9}$ and $\mu = 2$,
- the MTS-BPX preconditioner $\hat{\mathfrak{M}}_k$, denoted by MTS-BPX,
- the diagonal preconditioner $\mathrm{diag}[\mathfrak{v}]$, where $\mathfrak{v}$ is the main diagonal of $A_{\mathcal{R}_2}$.

For polynomial degrees $p < 100$, the multiple-tridiagonal scaling BPX preconditioner $\hat{\mathfrak{M}}_k$ (6.2.3) is the fastest method in order to solve (6.1). For polynomial degrees $p > 100$, the preconditioner $\breve{\mathfrak{M}}_{k,S_{2,k},1}$ beats the MTS-BPX preconditioner. However, these two preconditioners and the preconditioners $\bar{\mathfrak{M}}_{k,S_{1,k},1}$ and $\bar{\mathfrak{M}}_{k,S_{2,k},1}$ lie in a relatively small time range, e.g. for $p = 1023$ between 111 and 147 seconds, cf. Tables 6.1, 6.2 and 6.4. The two preconditioners, for which Theorem 6.1 holds, the multi-grid preconditioner $\bar{\mathfrak{M}}_{k,S_{0,k},3}$ and the AMLI preconditioner $\tilde{\mathfrak{M}}_{k,\frac{17}{9},2}$ need about twice as many time as the other ones.

7 Future work-wavelets

In chapter 4, we have considered finite element and finite difference discretizations for several problems in one, two and three space dimensions. Most of the discretizations in 2D and 3D are tensor-product discretizations of corresponding problems in one dimension.
In this chapter, we will derive wavelet preconditioners for the solution of the corresponding systems of linear algebraic equations. We will give only some ideas, the condition number estimates will be proved in the future. Moreover, we will propose preconditioners for the element stiffness matrices of the p-version of the fem, $A_{\mathcal{R}_2}$ and $A_{\mathcal{R}_3}$ (3.3.4).

7.1 1D case, motivation

We consider problem (4.1.3): Find $u \in H_0^1((0,1)) \cap L^2_\omega((0,1)) \cap L^2_{\omega^{-1}}((0,1))$ such that

$$a_1(u,v) = \int_0^1 (u'(x)v'(x) + \omega^2(x)u(x)v(x) + \omega^{-2}u(x)v(x))\,\mathrm{d}x = \langle g, v\rangle \tag{7.1.1}$$

holds for all $v \in H_0^1((0,1)) \cap L^2_\omega((0,1)) \cap L^2_{\omega^{-1}}((0,1))$. The weight function $\omega^2(x)$ is specified later. As described in subsection 4.1.2, we discretize problem (7.1.1) by piecewise linear elements on the mesh $T_k = \bigcup_{i=0}^{n-1} \left(\frac{i}{n}, \frac{i+1}{n}\right)$, where $n = 2^k$ and k denotes the level number. Let $\{\phi_i^{(1,k)}\}_{i=1}^{n-1}$ be the basis of the usual hat functions (4.1.4). We introduce the matrices

$$\begin{aligned} M_\omega^\phi &= \left[\langle \phi_j^{(1,k)}, \phi_i^{(1,k)}\rangle_\omega\right]_{i,j=1}^{n-1}, \\ T_{\omega=1}^\phi &= \left[\langle (\phi_j^{(1,k)})', (\phi_i^{(1,k)})'\rangle_{\omega=1}\right]_{i,j=1}^{n-1}, \end{aligned}$$

where $\langle\cdot,\cdot\rangle_\omega$ denotes the $L^2_\omega((0,1))$ scalar product, i.e.

$$\langle u, v\rangle_\omega = \int_0^1 \omega^2(x)u(x)v(x)\,\mathrm{d}x.$$

In subsection 4.1.2, see (4.1.6), (4.1.8) and (4.1.7), we have shown that $T_{\omega=1}^\phi = 2nT_2$, $M_{\omega=x}^\phi = \frac{1}{6n^3}D_5$, and $M_{\omega=1/x}^\phi = 4nD_6$. The matrices T_2, D_5 and D_6 are defined via relations (3.4.7), (3.4.9) and (3.4.10). Moreover, the matrices D_5 and D_6 corresponding to the mass parts $\langle\cdot,\cdot\rangle_\omega$ and $\langle\cdot,\cdot\rangle_{\omega^{-1}}$ of the bilinear form $a_1(\cdot,\cdot)$ (7.1.1) are spectrally equivalent to the diagonal matrix D_3 (3.4.1) and its inverse D_3^{-1}, cf. Lemma 3.8 and Lemma 3.9. However, for the matrix $T_2 \in \mathbb{R}^{n-1\times n-1}$ corresponding to the stiffness part in the bilinear form $a_1(\cdot,\cdot)$, it is not known

a diagonal matrix $D \in \mathbb{R}^{n-1\times n-1}$ such that the condition number of $D^{-1}T_2$ is bounded by a constant independent of the dimension $n-1$. Let $\{\phi_i^{(1,l)}\}_{(i,l)\in\hat{I}_k}$ be the hierarchical basis, see [80], on level k. The index set $\hat{I}_k$ is given by

$$\hat{I}_k = \left\{(i,l) \in \mathbb{N}^2, 1 \le l \le k, i = 2m-1, 1 \le m \le 2^{l-1}, m \in \mathbb{N}\right\}.$$

Let

$$T^{\phi,h}_{\omega=1} = \left[\langle(\phi_j^{(1,l')})', (\phi_i^{(1,l)})'\rangle_{\omega=1}\right]_{(i,l),(j,l')\in\hat{I}_k}$$

be the matrix corresponding to the stiffness part of the bilinear form (7.1.1) with respect to the hierarchical basis $\{\phi_i^{(1,l)}\}_{(i,l)\in\hat{I}_k}$. Then, by a simple calculation, the matrix $T^{\phi,h}_{\omega=1}$ is a diagonal matrix. More precisely, one obtains

$$\langle(\phi_j^{(1,l')})', (\phi_i^{(1,l)})'\rangle_{\omega=1} = 2^l\delta_{ll'}\delta_{ij}.$$

Thus, we have found a basis in which the stiffness part of the bilinear form $a_1(\cdot,\cdot)$ is spectrally equivalent to a diagonal matrix. However, a diagonal matrix D is not known such that the mass matrix

$$M^{\phi,h}_{\omega} = \left[\langle\phi_j^{(1,l')}, \phi_i^{(1,l)}\rangle_{\omega}\right]_{(i,l),(j,l')\in\hat{I}_k}$$

with respect to the hierarchical basis satisfies the condition number estimate $\kappa\left(D^{-1}M^{\phi,h}_{\omega}\right) < c$ independent of the dimension of the matrices.
Consider (7.1.1) with the weight function $\omega(x) = 1$. In the wavelet theory, see e.g. [29], [71], it is known that it can be constructed a basis $\{\psi_j^l\}_{l\le k}$ with $\operatorname{span}\{\psi_j^l\}_{l\le k} = \operatorname{span}\{\phi_i^{(1,k)}\}_{i=1}^{n-1}$ such that the matrices

$$\begin{aligned} M^{\psi}_{\omega=1} &= \left[\langle\psi_{j'}^{l'}, \psi_j^l\rangle_{\omega=1}\right]_{(j,l),(j',l')} \quad \text{and} \\ T^{\psi}_{\omega=1} &= \left[\langle(\psi_{j'}^{l'})', (\psi_j^l)'\rangle_{\omega=1}\right]_{(j,l),(j',l')} \end{aligned}$$

are spectrally equivalent to diagonal matrices. More precisely, let $D_{M^{\psi}_{\omega=1}}$ be the identity matrix I and $D_{T^{\psi}_{\omega=1}} = \operatorname{diag}[\mathfrak{u}]$, where $\mathfrak{u} = \left[2^{2l}\right]_{(j,l)}$. Then, see [29], [71], there

$$\kappa\left((D_{M^{\psi}_{\omega=1}})^{-1}M^{\psi}_{\omega=1}\right) = \mathcal{O}(1), \tag{7.1.2}$$

$$\kappa\left((D_{T^{\psi}_{\omega=1}})^{-1}T^{\psi}_{\omega=1}\right) = \mathcal{O}(1) \tag{7.1.3}$$

holds. These facts can be used to derive a preconditioner for $T^{\phi}_{\omega=1}$ and $M^{\phi}_{\omega=1}$. Let Q be the basis transformation from the nodal basis $\{\phi_i^{(1,k)}\}_{i=1}^{2^k-1}$ to the wavelet basis $\{\psi_j^l\}_{l\le k}$. Then,

$$T^{\psi}_{\omega=1} = Q^T T^{\phi}_{\omega=1} Q.$$

By $\kappa\left((D_{T^{\psi}_{\omega=1}})^{-1}T^{\psi}_{\omega=1}\right) = \mathcal{O}(1)$, the condition number estimates

$$\kappa\left((D_{T^{\psi}_{\omega=1}})^{-1}Q^T T^{\phi}_{\omega=1}Q\right) = \mathcal{O}(1) \quad \Longleftrightarrow \quad \kappa\left(Q(D_{T^{\psi}_{\omega=1}})^{-1}Q^T T^{\phi}_{\omega=1}\right) = \mathcal{O}(1)$$

are valid. Similarly, $\kappa\left(Q(D_{M^{\psi}_{\omega=1}})^{-1}Q^T M^{\phi}_{\omega=1}\right) = \mathcal{O}(1)$ is valid. Thus, we have found preconditioners for $T^{\phi}_{\omega=1}$, and $M^{\phi}_{\omega=1}$.
In the case of the singular weight functions $\omega(x) = x$ and $\omega(x) = \frac{1}{x}$, a result of the type $\kappa\left(Q(D_{M^{\psi}_{\omega}})^{-1}Q^T M^{\phi}_{\omega}\right) = \mathcal{O}(1)$ is not known for a wavelet basis $\{\psi^l_j\}_{l\leq k}$. This result will be shown in the future work. Because of the importance, we add the results here. We will formulate the corresponding theorem only.

THEOREM 7.1. *It exists a wavelet basis $\{\psi^l_j\}_{l\leq k} \subset \mathbb{V}_k$ such that the following assertions hold:*

- *The matrix $T^{\psi}_{\omega=1}$ is spectrally equivalent to the matrix $D_{T^{\psi}_{\omega=1}} = \mathrm{diag}[\mathfrak{b}]$, where $\mathfrak{b} = \left[2^{2l}\right]^T_{(j,l)}$, i.e. $\kappa\left((D_{T^{\psi}_{\omega=1}})^{-1}T^{\psi}_{\omega=1}\right) = \mathcal{O}(1)$.*
- *The matrix M^{ψ}_{ω} is spectrally equivalent to the matrix $D_{M^{\psi}_{\omega}} = \mathrm{diag}[\mathfrak{t}]$, where $\mathfrak{t} = \left[\omega^2(2^{-l}j)\right]^T_{(j,l)}$, i.e. $\kappa\left((D_{M^{\psi}_{\omega}})^{-1}M^{\psi}_{\omega}\right) = \mathcal{O}(1)$.*

Proof: The proof will be given in a forthcoming paper together with Reinhold Schneider and Christoph Schwab.□

7.2 2D and 3D case

Using a wavelets basis $\{\psi^l_j\}_{l\leq k}$, preconditioners can be derived for the systems of linear algebraic equations arising from the discretizations of (4.2.10), (4.2.14), (4.3.1), and (4.3.2). We explain the idea in the case of problem (4.2.14) with the bilinear form

$$a_2(u,v) = \int_{\Omega} 2\omega^2(x)u_y v_y + 2\omega^2(y)u_x v_x + \left(\frac{\omega^2(x)}{\omega^2(y)} + \frac{\omega^2(y)}{\omega^2(x)}\right)uv.$$

For the problems (4.2.10), (4.3.1), and (4.3.2), it can be done by the same arguments. The discretization of (4.2.14) by piecewise bilinear finite elements on the mesh $\mathcal{E}^k_{ij}$ yields to a system of linear algebraic equations of the type

$$\begin{aligned} C_2\underline{u} &= \left((T_2 + D_6)\otimes D_5 + D_5 \otimes (T_2 + D_6)\right)\underline{u}, \\ &= c\left((2T^{\phi}_{\omega=1} + M^{\phi}_{\omega=x^{-1}})\otimes M^{\phi}_{\omega=x} + M^{\phi}_{\omega=x}\otimes(2T^{\phi}_{\omega=1} + M^{\phi}_{\omega=x^{-1}})\right)\underline{u} = \underline{g}, \end{aligned} \tag{7.2.1}$$

cf. Lemma 4.5. For each of the involved matrices, $T^{\phi}_{\omega=1}$, $M^{\phi}_{\omega=x^{-1}}$ and $M^{\phi}_{\omega=x}$, we propose a preconditioner of the type $Q^{-T}\hat{D}Q^{-1}$, where $\hat{D}$ is a properly chosen diagonal matrix. More precisely, we choose

- for $T^{\phi}_{\omega=1}$: $Q^{-T}D_{T^{\psi}_{\omega=1}}Q^{-1}$,
- for $M^{\phi}_{\omega=x}$: $Q^{-T}D_{M^{\psi}_{\omega=x}}Q^{-1}$,

- for $M^{\phi}_{\omega=x^{-1}}$: $Q^{-T} D_{M^{\psi}_{\omega=x^{-1}}} Q^{-1}$.

Therefore, cf. the properties of the Kronecker product in Lemma 2.5, the matrix

$$\begin{aligned} C_2^{\psi} &= \left(Q^{-T} \otimes Q^{-T}\right) \Big((2D_{T^{\psi}_{\omega=1}} + D_{M^{\psi}_{\omega=x^{-1}}}) \otimes D_{M^{\psi}_{\omega=x}} + \\ &\quad D_{M^{\psi}_{\omega=x}} \otimes (2D_{T^{\psi}_{\omega=1}} + D_{M^{\psi}_{\omega=x^{-1}}}) \Big) \left(Q^{-1} \otimes Q^{-1}\right) \end{aligned} \tag{7.2.2}$$

is the preconditioner for C_2, see (3.4.16). Since

$$D_2^{\psi} = (D_{T^{\psi}_{\omega=1}} + D_{M^{\psi}_{\omega=x^{-1}}}) \otimes D_{M^{\psi}_{\omega=x}} + D_{M^{\psi}_{\omega=x}} \otimes (D_{T^{\psi}_{\omega=1}} + D_{M^{\psi}_{\omega=x^{-1}}})$$

is a diagonal matrix, the inverse of C_2^{ψ} can easily be computed, i.e.

$$(C_2^{\psi})^{-1} = (Q \otimes Q) (D_2^{\psi})^{-1} \left(Q^T \otimes Q^T\right). \tag{7.2.3}$$

The matrix Q denotes the one dimensional fast wavelet transformation, the cost for $Q\underline{r}_1$ is $\mathcal{O}(n)$. Thus, the total cost for the multiplication $(C_2^{\psi})^{-1}\underline{r}$ is arithmetically optimal, i.e. $\mathcal{O}(n^2)$. In the same way, we define the wavelet preconditioners C_5^{ψ}, C_8^{ψ}, and C_9^{ψ} given by their inverses,

$$(C_5^{\psi})^{-1} = (Q \otimes Q) (D_5^{\psi})^{-1} \left(Q^T \otimes Q^T\right), \tag{7.2.4}$$

$$(C_8^{\psi})^{-1} = (Q \otimes Q \otimes Q) (D_8^{\psi})^{-1} \left(Q^T \otimes Q^T \otimes Q^T\right), \tag{7.2.5}$$

$$(C_9^{\psi})^{-1} = (Q \otimes Q \otimes Q) (D_9^{\psi})^{-1} \left(Q^T \otimes Q^T \otimes Q^T\right) \tag{7.2.6}$$

for C_5 (3.4.19), C_8 (3.4.22) and C_9 (3.4.23). The matrices D_5^{ψ}, D_8^{ψ} and D_9^{ψ} are the diagonal matrices

$$\begin{aligned} D_5^{\psi} &= D_{T^{\psi}_{\omega=1}} \otimes D_{M^{\psi}_{\omega=x}} + D_{M^{\psi}_{\omega=x}} \otimes D_{T^{\psi}_{\omega=1}}, \\ D_8^{\psi} &= D_{T^{\psi}_{\omega=1}} \otimes D_{T^{\psi}_{\omega=1}} \otimes D_{M^{\psi}_{\omega=x}} + D_{T^{\psi}_{\omega=1}} \otimes D_{M^{\psi}_{\omega=x}} \otimes D_{T^{\psi}_{\omega=1}} \\ &\quad + D_{M^{\psi}_{\omega=x}} \otimes D_{T^{\psi}_{\omega=1}} \otimes D_{T^{\psi}_{\omega=1}}, \\ D_9^{\psi} &= (2D_{T^{\psi}_{\omega=1}} + D_{M^{\psi}_{\omega=x^{-1}}}) \otimes (2D_{T^{\psi}_{\omega=1}} + D_{M^{\psi}_{\omega=x^{-1}}}) \otimes D_{M^{\psi}_{\omega=x}} \\ &\quad + (2D_{T^{\psi}_{\omega=1}} + D_{M^{\psi}_{\omega=x^{-1}}}) \otimes D_{M^{\psi}_{\omega=x}} \otimes (2D_{T^{\psi}_{\omega=1}} + D_{M^{\psi}_{\omega=x^{-1}}}) \\ &\quad + D_{M^{\psi}_{\omega=x}} \otimes (2D_{T^{\psi}_{\omega=1}} + D_{M^{\psi}_{\omega=x^{-1}}}) \otimes (2D_{T^{\psi}_{\omega=1}} + D_{M^{\psi}_{\omega=x^{-1}}}). \end{aligned}$$

REMARK 7.2. *Using Theorem 7.1, there* $\kappa\left((C_i^{\psi})^{-1} C_i\right) = \mathcal{O}(1)$ *holds for* $i = 2, 5, 8, 9$.

7.3 Example of a wavelet basis

In this section, a wavelet basis is given which satisfies Theorem 7.1 in the case of the weight function $\omega(x) = 1$. We refer to the papers [25] and [26] for the construction of such a multi-resolution basis. The so called mother wavelet is a linear combination of the nodal hat functions

$\phi_j^{(1,k)}$, $j = 1\ldots,5$, see (4.1.4), i.e.

$$\begin{aligned}\psi_1^3(x) &= -\frac{1}{8}\phi_1^{(1,3)}(x) - \frac{1}{4}\phi_2^{(1,3)}(x) + \frac{3}{4}\phi_3^{(1,3)}(x) - \frac{1}{4}\phi_4^{(1,3)}(x) - \frac{1}{8}\phi_5^{(1,3)}(x) \qquad (7.3.1)\\ &= \begin{cases} -x & \text{if} \quad x \in [0, \frac{1}{4}] \\ 8x - \frac{9}{4} & \text{if} \quad x \in [\frac{1}{4}, \frac{3}{8}] \\ -8x + \frac{15}{4} & \text{if} \quad x \in [\frac{3}{8}, \frac{1}{2}] \\ x - \frac{3}{4} & \text{if} \quad x \in [\frac{1}{2}, \frac{3}{4}] \\ 0 & \text{else} \end{cases}.\end{aligned}$$

In the wavelet literature [71], this wavelet is denoted as ψ_{22} because it has two vanishing moments on the primal and dual side. The family of wavelets $\{\psi_j^l\}$ are constructed via translations and compressions. More precisely, let

$$\psi_j^l = 2^{\frac{l}{2}}\psi_1^3\left(\frac{1}{8}(2^l x - 2(j-1))\right) \quad 1 \le j \le 2^{l-2}, 3 \le l \le k, j, l \in \mathbb{N}. \qquad (7.3.2)$$

Figure 7.1 displays one wavelet of the family $\{\psi_j^l\}$. On the boundary at $x = 0$, we define, [25],

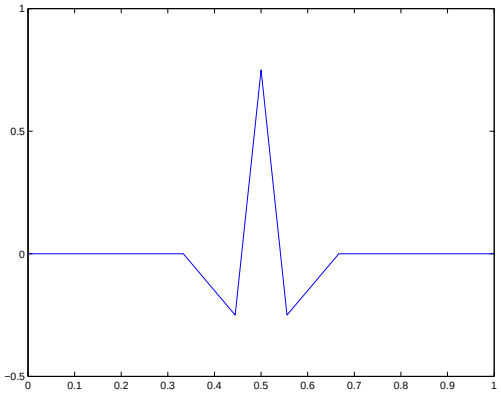

Figure 7.1: Wavelet ψ_j^l.

$$\begin{aligned}\psi_0^3(x) &= \frac{7}{8}\phi_1^{(1,3)}(x) - \frac{1}{4}\phi_2^{(1,3)}(x) - \frac{1}{8}\phi_3^{(1,3)}(x) \qquad (7.3.3)\\ &= \begin{cases} 7x & \text{if} \quad x \in [0, \frac{1}{8}] \\ -9x + 2 & \text{if} \quad x \in [\frac{1}{8}, \frac{1}{4}] \\ x - \frac{1}{2} & \text{if} \quad x \in [\frac{1}{4}, \frac{1}{2}] \\ 0 & \text{else} \end{cases}\end{aligned}$$

and $\psi_0^l(x) = 2^{\frac{l}{2}}\psi_0^3(2^{l-3}x)$. On the boundary at $x = 1$, let

$$\psi_{2^{l-2}+1}^l(x) = \psi_0^l(1-x) \quad \text{for} \quad l \ge 3.$$

Moreover, let by definition

$$\psi_j^2(x) \quad = \quad \phi_{j+1}^{(1,2)}(x) \quad \text{for} \quad j = 0, 1, 2.$$

Now, the system of wavelet functions $\left\{\psi_j^k\right\}_{l=2,j=0}^{k,2^{l-2}+1}$ is a basis in $\mathbb{V}_k$.

7.4 Application to the p-version and numerical experiments

Similarly as the multi-grid preconditioners in chapter 6, we can use the wavelet preconditioners C_2^ψ, C_5^ψ, C_8^ψ, C_9^ψ, (7.2.3)-(7.2.6) as preconditioner for the p-version element stiffness matrices $A_{\mathcal{R}_2}$ and $A_{\mathcal{R}_3}$ (3.3.4). For $A_{\mathcal{R}_2}$, we define the preconditioners

$$\mathcal{W}_i = P^T \text{blockdiag}\left[C_i^\psi\right]_{j=1}^4 P \quad \text{for} \quad i = 2, 5 \tag{7.4.1}$$

with the permutation matrix P of Proposition 3.3. For $A_{\mathcal{R}_3}$, let

$$\mathcal{W}_i = \hat{P}^T \text{blockdiag}\left[C_i^\psi\right]_{j=1}^8 \hat{P} \quad \text{for} \quad i = 8, 9 \tag{7.4.2}$$

be the preconditioners. The matrix $\hat{P}$ denotes the permutation matrix $\hat{P}$ of Proposition 3.4.

THEOREM 7.3. *The following condition number estimates are valid:*

- $\kappa\left({\mathcal{W}_5}^{-1} A_{\mathcal{R}_2}\right) \leq c(1 + \log p)$,
- $\kappa\left({\mathcal{W}_2}^{-1} A_{\mathcal{R}_2}\right) \leq c$,
- $\kappa\left({\mathcal{W}_8}^{-1} A_{\mathcal{R}_3}\right) \leq c(1 + \log p)^2$,
- $\kappa\left({\mathcal{W}_9}^{-1} A_{\mathcal{R}_3}\right) \leq c$.

The parameter c denotes a constant which is independent of the polynomial degree p.

Proof: The result follows from Propositions 3.3, 3.4, Theorems 3.11 and 3.12, Theorem 7.1 and Remark 7.2.□
Now, we give some numerical examples. All calculations are done on a Pentium III, 800 Mhz. The systems of linear algebraic equations

$$A_{\mathcal{R}_2}\underline{u} \quad = \quad \underline{f}, \tag{7.4.3}$$
$$A_{\mathcal{R}_3}\underline{u} \quad = \quad \underline{f} \tag{7.4.4}$$

are solved using the preconditioned conjugate gradient method. In all numerical experiments, it is chosen a relative accuracy of $\varepsilon = 10^{-9}$ in the preconditioned energy norm. The preconditioners

$\mathcal{W}_5$ and $\mathcal{W}_2$ (7.4.1) are chosen as preconditioner for $A_{\mathcal{R}_2}$. For $A_{\mathcal{R}_3}$, we apply the preconditioners $\mathcal{W}_8$ and $\mathcal{W}_9$ (7.4.2). The corresponding wavelets are the wavelets $\{\psi_j^l\}_{l=2,j=0}^{k,2^{2^{l-2}+1}}$ defined via relations (7.3.1) and (7.3.3). Table 7.1 displays the numbers of iterations of the pcg-method and the time reducing the error up to a factor of $\varepsilon = 10^{-9}$ in order to solve (7.4.3) with the preconditioners $\mathcal{W}_5$ and $\mathcal{W}_2$. The numbers of iterations of the pcg-method in order to solve (7.4.4) with the

p	$\mathcal{W}_5$		$\mathcal{W}_2$
	It	time [sec]	It
3	3	0.001	3
7	22	0.002	23
15	30	0.010	31
31	36	0.044	36
63	40	0.192	41
127	46	1.066	45
255	50	5.34	49
511	55	24.01	54
1023	58	120.6	57

Table 7.1: Numbers of iterations of the pcg-method for (7.4.3) using the preconditioners $\mathcal{W}_5$ and $\mathcal{W}_2$.

preconditioners $\mathcal{W}_8$ and $\mathcal{W}_9$ are displayed on Table 7.2. The numbers of iterations do not differ

p	$\mathcal{W}_8$	$\mathcal{W}_9$
3	3	3
7	41	43
15	50	52
31	56	57
63	64	63
127	74	70

Table 7.2: Numbers of iterations of the pcg-method for (7.4.4) using the preconditioners $\mathcal{W}_8$ and $\mathcal{W}_9$.

significantelly between $\mathcal{W}_5$ and $\mathcal{W}_2$, and, $\mathcal{W}_8$ and $\mathcal{W}_9$. In all cases, the numbers of iterations grow slightly. In comparison to the most multi-grid preconditioners $\bar{\mathfrak{M}}_{k,S,\mu}$ (6.1.2) and $\breve{\mathfrak{M}}_{k,S,\mu}$ (6.2.2), and the AMLI preconditioners $\tilde{\mathfrak{M}}_{k,r,\mu}$ (6.1.1), of chapter 6, the total numbers of iterations of the pcg-method are relatively high for the wavelet preconditioners $\mathcal{W}_5$ and $\mathcal{W}_2$. However, the cost in order to apply $\mathcal{W}_i^{-1}\underline{r}$, $i = 2, 5$ is cheaper than the cost for the multi-grid preconditioning operation $(\bar{\mathfrak{M}}_{k,S,\mu})^{-1}\underline{r}$, or $(\breve{\mathfrak{M}}_{k,S,\mu})^{-1}\underline{r}$. So, the time in order to reduce the error up to a factor of

10^{-9} is as good as for the fastest multi-level preconditioners like the MTS-BPX preconditioner $\hat{\mathfrak{M}}_k$ (6.2.3).

Remarks to the estimate of the strengthened Cauchy-inequality

Here, we give the exact values for the parameters p (5.3.27) and q (5.3.28). We set

$$\begin{aligned} r &= i-1, \\ s &= j-1. \end{aligned}$$

Then, we obtain the following results for p and q.

$$\begin{aligned} p := \frac{1}{704}(&5857266360\,s + 4407665790\,r + 1508755050 + 146252736\,s^6 \\ &+ 1111426560\,s^5 + 27808704\,r^6 + 302620032\,r^5 + 9324984713\,s^2 \\ &+ 5434977449\,r^2 + 3923127840\,s^4 + 7936810608\,s^3 \\ &+ 3647255568\,r^3 + 1415409600\,r^4 + 9269249088\,s^4\,r \\ &+ 8601027360\,s^4\,r^2 + 20130620928\,s^3\,r + 20920075392\,s^3\,r^2 \\ &+ 17559686400\,s^2\,r^3 + 6376566048\,s^2\,r^4 + 12919365888\,s\,r^3 \\ &+ 4918733952\,s\,r^4 + 124830720\,s^2\,r^6 + 1326974976\,s^2\,r^5 \\ &+ 3982219776\,s^4\,r^3 + 3786647040\,s^3\,r^4 + 11609339904\,s^3\,r^3 \\ &+ 277115904\,s^6\,r^2 + 328872960\,s^6\,r + 2493112320\,s^5\,r \\ &+ 999364608\,s^4\,r^4 + 2094465024\,s^5\,r^2 + 151621632\,s^4\,r^5 \\ &+ 735657984\,s^3\,r^5 + 69672960\,s^3\,r^6 + 108158976\,s^5\,r^4 \\ &+ 779452416\,s^5\,r^3 + 14432256\,s^4\,r^6 + 14432256\,s^6\,r^4 \\ &+ 103514112\,s^6\,r^3 + 1047619584\,s\,r^5 + 97625088\,s\,r^6 \\ &+ 28493849120\,s^2\,r^2 + 25194885712\,s^2\,r + 19809599216\,s\,r^2 \\ &+ 16586949280\,s\,r)/((20\,r + 17 + 6\,r^2)(82016\,s + 76846\,r \\ &+ 65589\,s^2 + 58245\,r^2 + 47232\,s^2\,r^2 + 93456\,s^2\,r + 93168\,s\,r^2 \\ &+ 139936\,s\,r + 4896\,s^4 + 26112\,s^3 + 21120\,r^3 + 3168\,r^4 \\ &+ 5760\,s^4\,r + 1728\,s^4\,r^2 + 30720\,s^3\,r + 9216\,s^3\,r^2 + 11520\,s^2\,r^3 \\ &+ 1728\,s^2\,r^4 + 30720\,s\,r^3 + 4608\,s\,r^4 + 39930)(6\,s^2 + 16\,s + 11)) \end{aligned}$$

$$
\begin{aligned}
q := \frac{1}{123904}(&3175524000\,s + 10752404850\,r + 925888320\,s^6 \\
&+ 3527193600\,s^5 + 153679680\,r^6 + 2180787840\,r^5 \\
&+ 6123829635\,s^2 + 25829259555\,r^2 + 5339341800\,s^4 \\
&+ 5845588560\,s^3 + 24034055760\,r^3 + 10651944600\,r^4 \\
&+ 18162835680\,s^4\,r + 24937019664\,s^4\,r^2 + 42653867520\,s^3\,r \\
&+ 81996584832\,s^3\,r^2 + 120359893824\,s^2\,r^3 + 52045531152\,s^2\,r^4 \\
&+ 90435290880\,s\,r^3 + 39407913600\,s\,r^4 + 742404096\,s^2\,r^6 \\
&+ 10535067648\,s^2\,r^5 + 17602460928\,s^4\,r^3 + 29858095872\,s^3\,r^4 \\
&+ 70165140480\,s^3\,r^3 + 1735243776\,s^6\,r^2 + 2071802880\,s^6\,r \\
&+ 7892582400\,s^5\,r + 6669527040\,s^4\,r^4 + 6610452480\,s^5\,r^2 \\
&+ 1260582912\,s^4\,r^5 + 5998067712\,s^3\,r^5 + 422682624\,s^3\,r^6 \\
&+ 338411520\,s^5\,r^4 + 2450718720\,s^5\,r^3 + 88833024\,s^4\,r^6 \\
&+ 88833024\,s^6\,r^4 + 643313664\,s^6\,r^3 + 8005662720\,s\,r^5 \\
&+ 564157440\,s\,r^6 + 136254292064\,s^2\,r^2 + 65646211760\,s^2\,r \\
&+ 100770474640\,s\,r^2 + 46577704800\,s\,r)/((20\,r + 17 + 6\,r^2)(\\
82016\,s &+ 76846\,r + 65589\,s^2 + 58245\,r^2 + 47232\,s^2\,r^2 \\
&+ 93456\,s^2\,r + 93168\,s\,r^2 + 139936\,s\,r + 4896\,s^4 + 26112\,s^3 \\
&+ 21120\,r^3 + 3168\,r^4 + 5760\,s^4\,r + 1728\,s^4\,r^2 + 30720\,s^3\,r \\
&+ 9216\,s^3\,r^2 + 11520\,s^2\,r^3 + 1728\,s^2\,r^4 + 30720\,s\,r^3 + 4608\,s\,r^4 \\
&+ 39930)(6\,s^2 + 16\,s + 11))
\end{aligned}
$$

Obviously, $p > 0$ and $q \geq 0$ hold for $i, j \geq 1$. Moreover, we can conclude

$$q = 0 \iff i = 1 \quad \text{and} \quad j = 1.$$

Hence, the estimate of Lemma 5.18 is sharp.

Bibliography

[1] M. Ainsworth. A preconditioner based on domain decomposition for h-p finite element approximation on quasi-uniform meshes. *SIAM J. Numer. Anal.*, 33(4):1358–1376, 1996.

[2] M. Ainsworth and B. Gao. An additive Schwarz preconditioner for p-version boundary element approximation of the hypersingular operator in three dimensions. *Numer. Math.*, 85(3):343–366, 2000.

[3] O. Axelsson. *Iterative Solution Methods*. Cambridge University Press, Cambridge, 1994.

[4] O. Axelsson and A. Padiy. On the additive version of the algebraic multilevel iteration for anisotropic elliptic problems. *SIAM J. Sci. Comp.*, 20(5):1807–1830, 1999.

[5] O. Axelsson and P.S. Vassilevski. Algebraic multilevel preconditioning methods I. *Numer. Math.*, 56:157–177, 1989.

[6] O. Axelsson and P.S. Vassilevski. Algebraic multilevel preconditioning methods II. *SIAM J. Numer. Anal.*, 27(6):1569–1590, 1990.

[7] I. Babuška, A. Craig, J. Mandel, and J. Pitkäranta. Efficent preconditioning for the p-version finite element method in two dimensions. *SIAM J.Numer.Anal.*, 28(3):624–661, 1991.

[8] I. Babuška, M. Griebel, and J. Pitkäranta. The problem of selecting the shape functions for a p-type finite element. *Int. Journ. Num. Meth. Eng.*, 28:1891–1908, 1989.

[9] S. Beuchler. Lösungsmethoden bei der p-version der FEM. Diplomarbeit, TU Chemnitz, December 1999.

[10] S. Beuchler. Lösungsmethoden bei der p-version der fem. Diplomarbeit, TU Chemnitz, December 1999.

[11] S. Beuchler. A preconditioner for solving the inner problem of the p-version of the FEM. Technical Report SFB393 00-25, Technische Universität Chemnitz, May 2000.

[12] S. Beuchler. The MTS-BPX-preconditioner for the p-version of the FEM. Technical Report SFB393 01-16, Technische Universität Chemnitz, May 2001.

[13] S. Beuchler. Multi-grid solver for the inner problem in domain decompostion methods for p-FEM. *SIAM J. Numer. Anal.*, 40(3):928–944, 2002.

[14] S. Börm and R. Hiptmair. Analysis of tensor product multigrid. *Numer. Algorithms*, 26(3):219–234, 2001.

[15] D. Braess. The contraction number of a multigrid method for solving the Poisson equation. *Numer. Math*, 37:387–404, 1981.

[16] D. Braess. *Finite Elemente*. Springer. Berlin-Göttingen-Heidelberg, 1991.

[17] J. Bramble, J. Pasciak, and A. Schatz. The construction of preconditioners for elliptic problems by substructuring I. *Math. Comp.*, 47(175):103–134, 1986.

[18] J. Bramble, J. Pasciak, and A. Schatz. The construction of preconditioners for elliptic problems by substructuring II. *Math. Comp.*, 49(179):1–16, 1987.

[19] J. Bramble, J. Pasciak, and A. Schatz:. The construction of preconditioners for elliptic problems by substructuring III. *Math. Comp.*, 51(184):415–430, 1988.

[20] J. Bramble, J. Pasciak, and A. Schatz. The construction of preconditioners for elliptic problems by substructuring IV. *Math. Comp.*, 53(187):1–24, 1989.

[21] J. Bramble, J. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55(191):1–22, 1991.

[22] J. Bramble and X. Zhang. Uniform convergence of the multigrid v-cycle for an anisotropic problem. *Math. Comp.*, 70(234):453–470, 2001.

[23] W. Cao and B. Guo. Preconditioning on element interfaces for the p-version finite element method and spectral element method. *SIAM J. Sci. Comput.*, 21(2):522–551, 1999.

[24] P. Ciarlet. *The Finite Element Method for Elliptic Problems*. North–Holland, Amsterdam, 1978.

[25] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harm. Anal.*, 1:54–81, 1993.

[26] A. Cohen, I. Daubechies, and P. Vial. Biorthogonal spline wavelets on the interval — Stability and moment conditions. *Appl. Comput. Harm. Anal.*, 6(2):132–196, 1998.

[27] D. Concus, G.H. Golub, and G. Meurant. Block preconditioning for the conjugate gradient method. *SIAM J.Sci.Stat.Comput.*, 6(1):220–252, 1985.

[28] J. W. Cooley and J. W. Tuckey. An algorithm for the machine calculation of complex fourier series. *Math. Comp.*, 19:297–301, 1965.

[29] W. Dahmen. Wavelet and multiscale methods for operator equations. *Acta Numerica*, 6:55–228, 1997.

[30] M. O. Deville and E. H. Mund. Finite element preconditioning for pseudospectral solutions of elliptic problems. *SIAM J. Sci. Stat. Comp.*, 18(2):311–342, 1990.

[31] F.R. Gantmacher. *Matrizenrechnung II*. Deutscher Verlag der Wissenschaften. Berlin, 1971.

[32] A. George. Nested dissection of a regular finite element mesh. *SIAM J. Numer. Anal.*, 10:345–363, 1973.

[33] A. George and J.W.-H. Liu. *Computer solution of large sparse positive definite systems*. Prenctice-Hall Inc. Englewood Cliffs. New Jersey, 1981.

[34] G.H. Golub and C.F. van Loan. *Matrix Computation*. The Johns Hopkins University Press. Baltimore-London, 1983.

[35] B. Guo and W. Cao. A preconditioner for the h-p version of the finite element method in two dimensions. *Numer. Math.*, 75:59–77, 1996.

[36] B. Guo and W. Cao. An iterative and parallel solver based on domain decomposition for the $h - p$ version of the finite element method. *J. Comput. Appl. Math.*, 83:71–85, 1997.

[37] B. Guo and E. P. Stephan. The h-p version of the coupling of finite element and boundary element methods for transmission problems in polyhedral domains. *Numer. Math.*, 80:87–107, 1998.

[38] G. Haase and U. Langer. Multigrid Methoden, 1998. Script zur Vorlesung, Johannes-Kepler-Universität Linz.

[39] G. Haase, U. Langer, and A. Meyer. The approximate Dirichlet domain decomposion method. part I: An algebraic approach. *Computing*, 47:137–151, 1991.

[40] W. Hackbusch. *Multigrid Methods and Applications*. Springer-Verlag. Heidelberg, 1985.

[41] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner Studienbücher Mathematik. Teubner–Verlag, Stuttgart, 1987.

[42] W. Hackbusch. *Iterative solution of Large Sparse Systems of Equations*. Number 95 in Applied Mathematical Sciences. Springer-Verlag. Berlin-Heidelberg-New York, 1993.

[43] W. Hackbusch and U. Trottenberg. *Multigrid Methods, Proceedings of the Conference held at Köln-Porz, November 23-27, 1981*. Number 960 in Lecture Notes in Mathematics. Springer-Verlag. Berlin-Heidelberg-New York, 1982.

[44] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49:409–436, 1952.

[45] N. Heuer and E.P. Stephan. Preconditioners for the p version of the Galerkin method for a coupled finite element/boundary element system. *Numer. Methods Partial Differ. Equations*, 14(1):49–61, 1998.

[46] N. Heuer, E.P. Stephan, and T.Tran. Multilevel additive schwarz method for the p version of the Galerkin boundary element method. *Math. Comput.*, 67(222):501–518, 1998.

[47] S.A. Ivanov and V.G. Korneev. On the preconditioning in the domain decomposition technique for the p-version finite element method. Part I. Technical Report SPC 95-35, Technische Universität Chemnitz-Zwickau, December 1995.

[48] S.A. Ivanov and V.G. Korneev. On the preconditioning in the domain decomposition technique for the p-version finite element method. Part II. Technical Report SPC 95-36, Technische Universität Chemnitz-Zwickau, December 1995.

[49] S. Jensen and V.G. Korneev. On domain decomposition preconditioning in the hierarchical $p-$version of the finite element method. *Comput. Methods. Appl. Mech. Eng.*, 150(1–4):215–238, 1997.

[50] M. Jung. Einige Klassen paraller iterativer Auflösungsverfahren. Habilitationsschrift, Technische Universität Chemnitz, 1999.

[51] M. Jung, U. Langer, A. Meyer, W. Queck, and M. Schneider. Multigrid preconditioners and their applications. Technical Report 03/89, Akad. Wiss. DDR, Karl-Weierstraß-Inst., 1989.

[52] G.M. Karniadakis and S.J.Sherwin. *Spectral/HP Element Methods for CFD*. Oxford University Press. Oxford, 1999.

[53] V. G. Korneev. Почти оптимальныи метод решения задач Дирихле на подобластях декомпосиции иерархическои hp-версии. Дифференциальные Уравнения, 37(7):1–15, 2001.

[54] A. Kufner and A.M. Sändig. *Some applications of weighted Sobolev spaces*. B.G.Teubner Verlagsgesellschaft. Leipzig, 1987.

[55] P. L. Lions. On the schwarz alternating method i. In R. Glowinski, G. H. Golub, G. A. Meurant, and J. P'eriaux, editors, *Proc. 1st Int. Symp. on Domain Decomposition Methods. SIAM, Philadelphia*, pages 1–42, Philadelphia, 1988. SIAM.

[56] P. L. Lions. On the schwarz alternating method ii. In *Proc. 1st Int. Symp. on Domain Decomposition Methods. Los Angeles*, pages 47–70, Los Angeles, 1989.

[57] J.F. Maitre and F. Musy. The contraction number of a class of two-level methods, and exact evaluation for some finite element subspaces and model problems. In W. Hackbusch and U. Trottenberg, editors, *Multigrid methods, Proceedings of the Conference held at Köln-Porz, November 23-27, 1981*, number 960 in Lecture Notes in Mathematics, pages 535–544, Berlin-Heidelberg-New York, 1982. Springer Verlag.

[58] J.F. Maitre and O. Pourquier. Condition number and diagonal preconditioning: Comparison of the p-version and the spectral element methods. *Numer. Math.*, 74(1):69–84, 1996.

[59] J. Mandel. An iterative solver for p-version finite elements in three dimensions. *Comput. Methods. Appl. Mech. Eng.*, 116:175–183, 1994.

[60] A. M. Matsokin and S. V. Nepomnyaschikh. The Schwarz alternation method in a subspace. *Iz. VUZ Mat.*, 29(10):61–66, 1985.

[61] A. M. Matsokin and S. V. Nepomnyaschikh. Norms in the space of traces of mesh functions. *Sov. J. Numer. Anal. Math. Modelling*, 3(3):199–216, 1988.

[62] G. Meurant. *Computer Solution of Large Linear Systems*, volume 28 of *Studies in Mathematics and its Applications*. Elsevier, Amsterdam, 1999.

[63] L. F. Pavarino. Additive schwarz methods for the p-version finite element method. *Numer. Math.*, 66(4):493–515, 1994.

[64] Th. Penzl. *Numerische Lösung großer Lyapunov-Gleichungen*. Logos-Verlag. Berlin, 1991.

[65] Ch. Pflaum. Fast and robust multilevel algorithms. Habilitationsschrift, Universität Würzburg, 1998.

[66] A. Quateroni and A. Valli. *Numerical Approximation of partial differential equations*. Number 23 in Springer Series in Computational Mathematics. Springer. Berlin-Heidelberg-New York, 1997.

[67] A. Reusken. A new lemma in multi-grid convergence theory. Technical Report RANA 91/07, Eindhoven, 1991.

[68] A. Reusken. On maximum norm convergence of multigrid methods for elliptic boundary value problems. *SIAM J. Numer. Anal.*, 31(2):378–392, 1994.

[69] A.A. Samarskij. *Theorie der Differenzenverfahren*. Akademische Verlagsgesellschaft Geest & Portig K.-G. Leipzig, 1984.

[70] N. Schieweck. A multi-grid convergence proof by a strengthened Cauchy-inequality for symmetric elliptic boundary value problems. In G. Telschow, editor, *Second multigrid seminar, Garzau 1985*, number 08-86 in Report R-Math, pages 49–62, Berlin, 1986. Karl-Weierstraß-Institut für Mathematik.

[71] R. Schneider. *Multiskalen- und Wavelet Matrixkompression*. Teubner, 1998.

[72] C. Schwab. *$p-$ and $hp-$finite element methods. Theory and applications in solid and fluid mechanics.* Clarendon Press. Oxford, 1998.

[73] H.R. Schwarz. *Numerische Mathematik.* B. G. Teubner. Stuttgart, 1993.

[74] G. Strang and G. Fix. *An Analysis of the Finite Element Method.* Prentice–Hall Inc., Englewood Cliffs, 1973.

[75] C.A. Thole. Beiträge zur Fourieranalyse von Mehrgittermethoden: V-cycle, ILU-Glättung, anisotrope Operatoren. Diplomarbeit, Universität Bonn, 1983.

[76] T. Tran and E. P. Stephan. Additive Schwarz algorithms for the p version of the galerkin boundary element method. *Numer. Math.*, 85(3):433–468, 2000.

[77] F.G. Tricomi. *Vorlesungen über Orthogonalreihen*. Springer. Berlin-Göttingen-Heidelberg, 1955.

[78] R. Verfürth. The contraction number of a multigrid method with mesh ratio 2 for solving Poisson's equation. *Lin. Algebra Appl.*, 60:113–128, 1984.

[79] A. J. Wathen. An analysis of some element-by-element techniques. *Comput. Methods Appl. Mech. Eng.*, 74(3):271–287, 1989.

[80] H. Yserentant. On the multi-level-splitting of the finite element spaces. *Numer. Math.*, 49:379–412, 1986.

[81] X. Zhang. Multilevel Schwarz methods. *Numer. Math.*, 63:521–539, 1992.

Theses

Multi-level methods for degenerated problems with applications to p-versions of the fem

Dipl.-Math. Sven Beuchler
Chemnitz University of Technology, Faculty of Mathematics

1. Computer simulations of many problems in natural and engineering sciences are based on the mathematical description of these problems by means of partial differential equations and appropriate boundary conditions. In most cases, these boundary value problems (bvp) cannot be solved analytically. A powerful tool to compute an approximate solution is the finite element method (fem). Mesh refinements or an increasing polynomial degree of the ansatz functions lead to an increasing accuracy of the approximate solution, if it is known that the exact solution of the bvp is sufficiently smooth. The first possibility is called h-version and the second one p-version of the fem. The combination of both gives hp-versions. As a result of the discretization one gets, in general, a large-scale system of algebraic equations

$$\mathcal{A}\underline{u} = \underline{f}\,. \tag{1}$$

   Usually, the matrix $\mathcal{A}$ is sparse. For symmetric, elliptic bvp's, the matrix $\mathcal{A}$ is symmetric and positive definite, but often ill-conditioned. Therefore, one needs appropriate preconditioners in order to get efficient solvers for the system of equations (1). In the theses, the construction of preconditioners for systems of finite element equations resulting from the p-version of the fem are discussed.

2. Most preconditioners for systems like (1) that arise from the discretization of bvp's by the p-version of the fem are based on domain decomposition (DD) techniques. For this purpose, we suppose that the considered domain is divided into q non-overlapping sub-domains. For two dimensional problems, the basis functions of the fem ansatz space are chosen in such a way that they can be divided into three groups:

   (vert) the vertex functions,

   (edg) the edge bubble functions,

   (int) the interior bubble functions.

   Analogously, the matrix $\mathcal{A}$ gets a block-structure:

$$\mathcal{A} \;=\; \begin{bmatrix} A_{vert} & A_{vert,edg} & A_{vert,int} \\ A_{edg,vert} & A_{edg} & A_{edg,int} \\ A_{int,vert} & A_{int,edg} & A_{int} \end{bmatrix}.$$

In a first step, one defines the preconditioner

$$C_p^{-1} = \begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I & \mathbf{0} \\ \mathbf{0} & -A_{int}^{-1}A_{int,edg} & I \end{bmatrix} \begin{bmatrix} A_{vert}^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & S^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & A_{int}^{-1} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I & -A_{edg,int}A_{int}^{-1} \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix},$$

where $S = A_{edg} - A_{edg,int}A_{int}^{-1}A_{int,edg}$ is the Schur complement. The condition number of $C_p^{-1}\mathcal{A}$ grows as $1 + \log p$, where p denotes the polynomial degree. The application of the preconditioner C_p requires the solution of systems of equations with the matrix A_{vert}, the Schur complement matrix S, and the sub-domain stiffness matrix A_{int}. In general, this is too expensive. Therefore, A_{vert}, S and A_{int} in the preconditioner C_p^{-1} are replaced by appropriate preconditioners. The matrix $-A_{int}^{-1}A_{int,edg}$ is replaced by an extension operator acting from the sub-domain boundaries into the interior of the sub-domains. Korneev and co-authors derived several preconditioners for the Schur complement S. The problem of the extension operator was discussed by Babuška et al. For A_{vert}, direct solvers or multi-grid methods can be applied.

3. The considered domain is the union of quadrilaterals. Each quadrilateral forms one sub-domain of the domain decomposition. Then, the matrix A_{int} is a block-diagonal matrix consisting of blocks $A_{int,i}$ which correspond to a particular sub-domain. A spectrally equivalent preconditioner for the matrix A_{int} is $C_{int} = \text{blockdiag}\,[A_{\mathcal{R}_2}]_{i=1}^{q}$. The matrix $A_{\mathcal{R}_2}$ is the element stiffness matrix related to the Dirichlet problem on the reference element $\mathcal{R}_2 = (-1,1)^2$. In the case of Poisson's equation, scaled integrated Legendre polynomials $\hat{L}_{ij}(x,y) = \hat{L}_i(x)\hat{L}_j(y)$, $2 \le i,j \le p$ are usually used for the basis of the interior functions. Then, the matrix $A_{\mathcal{R}_2}$ has about $5p^2$ nonzero elements and the condition number grows as p^2. For $A_{\mathcal{R}_2}$, we propose preconditioners of the type $C_{\mathcal{R}_2} = P\,\text{blockdiag}\,[K_k]_{i=1}^{4}\,P^T$, where the matrix K_k can be interpreted as a discretization matrix of a degenerated elliptic bvp using linear/bilinear finite elements or finite differences on uniform meshes or grids. Such degenerated problems are

$$-\omega^2(x)u_{yy} - \omega^2(y)u_{xx} = g \quad \text{or} \tag{2}$$

$$-\omega^2(x)u_{yy} - \omega^2(y)u_{xx} + 2\left(\frac{\omega^2(y)}{\omega^2(x)} + \frac{\omega^2(x)}{\omega^2(y)}\right)u = g \tag{3}$$

in $\Omega = (0,1)^2$, $u = 0$ on $\partial\Omega$, where $\omega(\xi) = \xi$. The matrix P is a suitably chosen permutation matrix. The condition number $\kappa\left({C_{\mathcal{R}_2}}^{-1}A_{\mathcal{R}_2}\right)$ grows as $1+\log p$ for (2), whereas the estimate $\kappa\left({C_{\mathcal{R}_2}}^{-1}A_{\mathcal{R}_2}\right) \le c$ is valid for (3).

4. Problem (2) with $\omega(\xi) = \xi$ is discretized by the h-version of the fem. A sequence of finite element discretizations with piecewise linear shape functions on uniform meshes T_l consisting of congruent, isosceles, right-angled triangles is investigated. The sequence of meshes $\{T_l\}_{l=1}^{k}$ is generated by a uniform refinement of the mesh T_1. The corresponding finite element spaces are denoted by $\mathbb{V}_l$ and can be split into the direct sum $\mathbb{V}_l = \mathbb{V}_{l-1} \bigoplus \mathbb{W}_l$, $l \ge 2$. A sequence of systems $\left\{K_l\underline{u}_l = \underline{g}_l\right\}_{l=1}^{k}$ arises as result of this discretization. A

multi-grid (k-grid) algorithm which can be interpreted as alternate, approximate projection onto the subspaces $\mathbb{V}_{l-1}$ and $\mathbb{W}_l$ is investigated. Therefore, systems with the matrix K_{l-1} and a matrix $K_{\mathbb{W}_l}$ have to be solved approximately. The matrix $K_{\mathbb{W}_l}$ is the stiffness matrix with respect to the new nodes on level l. The convergence rate σ_k of the considered multi-grid algorithm can be estimated purely algebraic. Firstly, it depends on the constant in the strengthened Cauchy-inequality and secondly on the convergence rate ρ_l, $l = 2, \ldots, k$ of the iterative solution procedure of $K_{\mathbb{W}_l}\underline{w} = \underline{r}$. For problem (2), an estimate of the constant of the strengthened Cauchy-inequality is derived. For the iterative solution of the system $K_{\mathbb{W}_l}\underline{w} = \underline{r}$, a special line smoother $S_{0,l}$ is built. Its error transion operator is given by $I - C_{\mathbb{W}_l}^{-1}K_{\mathbb{W}_l}$. Moreover, this construction is generalized to weight functions $\omega(\xi) = \xi^\alpha$ in (2), where $\alpha \geq 0$. The convergence rate of $S_{0,l}$ in order to solve $K_{\mathbb{W}_l}\underline{w} = \underline{r}$ is bounded by a constant $\rho < 1$. If the system $K_{l-1}\underline{u}_{l-1} = \underline{g}_{l-1}$, $l = 2, \ldots, k$, is solved by at least $\mu \geq 3$ iterations of the multi-grid algorithm for K_{l-1}, the convergence rate σ_k for the multi-grid algorithm satisfies the estimate $\sigma_k \leq \sigma < 1$. The arithmetical cost for one iteration of the multi-grid algorithm is proportional to the number of unknowns on the finest mesh T_k.

5. The ideas, which are used to define the matrix $C_{\mathbb{W}_l}$, can be transfered to the definition of a matrix R_l. Consequently, the matrix R_l corresponds to the space of all nodes on level l. If the unknowns are permuted, this matrix occurs as tridiagonal matrix. The smoother $S_{1,l}$, whose error transion operator is given by $S_{1,l} = I - \omega R_l^{-1}K_l$, can be interpreted as an ω-Jacobi-like smoother along the union of a horizontal and vertical line. The smoother $S_{1,l}$ operates on the whole approximation space. Numerical experiments indicate a multi-grid convergence rate $\sigma_k \leq \sigma < 1$ for a standard multi-grid algorithm with V-cycle ($\mu = 1$) and smoother $S_{1,l}$, $l = 2, \ldots, k$.

6. The multi-grid algorithms discussed in the theses 4 and 5 are used to define implicitly pre-conditioners $\overline{C}_{k,S,\mu}$. Here, S denotes the used smoother. The parameter μ is the number of iterations in order to solve the coarse grid problems. The matrix $\overline{C}_{k,S,\mu}$ is symmetric positive definite and the condition number of $\overline{C}_{k,S,\mu}^{-1}K_k$ is bounded by a constant independent of the mesh-size h for $\mu \geq 3$ and $S = S_{0,k}$. The application of the preconditioners $\overline{C}_{k,S,\mu}$ embedded in a preconditioned conjugate gradient method accelerates the convergence in comparison to the multi-grid algorithm applied to solve $K_k\underline{u}_k = \underline{f}_k$.

7. For the analysis of Algebraic Multi-level Iteration (AMLI) preconditioners $\tilde{C}_{k,\mu}$, it is assumed that the nodes are numbered hierarchically, i.e. first the nodes in the coarse mesh T_{l-1} and then the new ones in T_l, i.e. $K_{11} = K_{l-1}$ and $K_{22} = K_{\mathbb{W}_l}$. The AMLI preconditioner $\tilde{C}_{\mu,k}$ is recursively defined by

$$\tilde{C}_{l,\mu} = \begin{bmatrix} \tilde{C}^c_{l-1,\mu} & K_{12} + J_{12}(K_{22} - \tilde{C}_{22}) \\ \mathbf{0} & \tilde{C}_{22} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \tilde{C}_{22}^{-1}(K_{21} + (K_{22} - \tilde{C}_{22})J_{12}^T) & I \end{bmatrix},$$

with

$$(\tilde{C}^c_{l-1,\mu})^{-1} = (I - P_\mu(\tilde{C}^{-1}_{l-1,\mu}K_{l-1}))K_{l-1}^{-1}$$

for $l = 2, \ldots, k$ and $\tilde{C}_{1,\mu} = K_1$ for $l = 1$. The interpolation matrix J_{12} is defined in analogy to the interpolation matrix in the multi-grid algorithm. The matrix $\tilde{C}_{22}$ is a preconditioner for $K_{\mathbb{W}_l} = K_{22}$. For $\tilde{C}_{22}$, the matrix $\left(\lambda_{max}\left({C_{\mathbb{W}_l}}^{-1} K_{\mathbb{W}_l}\right)\right) C_{\mathbb{W}_l}$ is chosen. Taking a polynomial iteration with a Chebyshev polynomial P_μ of degree $\mu \geq 2$, the condition number of $(\tilde{C}_{\mu,k})^{-1} K_k$ is bounded by a constant independent of the mesh-size h.

8. In numerical experiments, the BPX preconditioner $\check{C}_k$ with multiple diagonal scaling for the matrix K_k shows a behaviour of $\kappa\left(\check{C}_k^{-1} K_k\right) \gg k^2$, where k denotes the level number. This behaviour can be improved by choosing a so called multiple tridiagonal scaling (MTS)-BPX preconditioner $\hat{C}_k$. In the case of the MTS-BPX preconditioner, a tridiagonal matrix R_l (see thesis 5) resulting from the smoother $S_{1,l}$ is used as scaling on each level $l = 2, \ldots, k$. Then, the upper eigenvalue estimate $\lambda_{max}\left(\hat{C}_k^{-1} K_k\right) \leq c(1+k)$ holds for weight functions of the type $\omega(\xi) = \xi^\alpha$ with $\alpha \geq 0$. Numerical experiments indicate that $\lambda_{min}\left(\hat{C}_k^{-1} K_k\right) \geq c$ and that the upper eigenvalue estimate is sharp.

9. The linear system $A_{\mathcal{R}_2} \underline{u} = \underline{f}$ (see thesis 3) can be solved in $\mathcal{O}(\sqrt{1+\log p})$ arithmetical operations by a preconditioned conjugate gradient method with the preconditioner $\mathcal{M}_k = P \,\mathrm{blockdiag}\left[M_k\right]_{i=1}^4 P^T$. The matrix M_k is a preconditioner for K_k and P is a permutation matrix. The condition number of $\mathcal{M}_k^{-1} A_{\mathcal{R}_2}$ is $\mathcal{O}(1+\log p)$ for the AMLI preconditioner $M_k = \tilde{C}_{\mu,k}$ (with $\mu \geq 2$) and the multi-grid preconditioner $M_k = \overline{C}_{k,\mu,S_{0,k}}$ (with $\mu \geq 3$). This estimate of the condition number is confirmed by numerical examples.

10. Wavelet preconditioners can be applied for systems arising from the fem-discretization of the one dimensional bvp $-u'' + \omega^2(x)u + \omega^{-2}(x)u = g$ in $(0,1)$ and $u(0) = u(1) = 0$ with piecewise linear elements on a uniform mesh. Preconditioners C_k^ψ for the corresponding tensor product problems in two and three dimensions are developed by tensor product arguments. These preconditioners C_k^ψ are used to derive preconditioners $\mathcal{W}_{d,j}$, $j = 1, 2$ for the p-version element stiffness matrices $A_{\mathcal{R}_d}$ of the reference element $\mathcal{R}_d = (-1,1)^d$ in two and three dimensions with $d = 2$ or $d = 3$, respectively. These preconditioners satisfy the condition number estimates $\kappa\left({\mathcal{W}_{d,1}}^{-1} A_{\mathcal{R}_d}\right) \leq \mathcal{O}(1+\log p)^{d-1}$, and $\kappa\left({\mathcal{W}_{d,2}}^{-1} A_{\mathcal{R}_d}\right) = \mathcal{O}(1)$.

Lebenslauf

Persönliche Daten

Name:	Sven Beuchler
Geburtsdatum:	24.Juli 1975
Geburtsort:	Karl-Marx-Stadt
Familienstand:	ledig

Schulausbildung

Sept.1982-Aug.1990	Polytechnische Oberschule in Karl-Marx-Stadt
Sept.1990-Aug.1992	Spezialschule math.-naturw.-techn. Richtung in Chemnitz
Sept.1992-Juni 1994	Johannes-Kepler-Gymnasium in Chemnitz Abschluß: Abitur

Zivildienst

Juli 1994-Sept.1995	Zivildienst

Studium

Okt.1995-Dez. 1999	Studium der Mathematik an der TU-Chemnitz; Abschluß als Diplom-Mathematiker

Wissenschaftlicher Werdegang

seit Jan. 2000	wissenschaftlicher Mitarbeiter an der Fakultät für Mathematik der TU-Chemnitz und im DFG-Sonderforschungsbereich 393 „Numerische Simulation auf massiv parallelen Rechnern”

Erklärung

Ich erkläre an Eides Statt, daß ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Chemnitz, den 5. Februar 2003